UNIVERSIDAD AUTÓNOMA DE MADRID
Departamento de Bioquímica

Doctoral Thesis

Identifying genomic biomarkers for cancer treatment

Sara Ruiz Pinto

Madrid, 2017

Departamento de Bioquímica
Facultad de Medicina
Universidad Autónoma de Madrid

Identifying genomic biomarkers for cancer treatment

Doctoral Thesis submitted by:

Sara Ruiz Pinto
M.Sc. in Molecular Biomedicine from Universidad Autónoma de Madrid in Madrid
B.Sc. in Biology from Universidad Complutense de Madrid in Madrid

Thesis director:

Dr. Anna González Neira

Human Genotyping-CeGen Unit
Human Cancer Genetics Programme
Spanish National Cancer Research Centre (CNIO)

Centro Nacional de Investigaciones Oncológicas

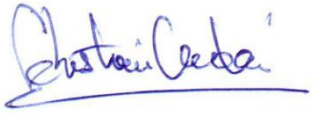
Dra. Anna González Neira, Jefa de la Unidad de Genotipado Humano-CeGen del Centro Nacional de Investigaciones Oncológicas (CNIO), como Directora

CERTIFICA:

Que **Sara Ruiz Pinto**, Licenciada en Biología por la Universidad Complutense de Madrid, ha realizado la presente Tesis Doctoral **"Identifying genomic biomarkers for cancer treatment"** y que a su juicio reúne plenamente todos los requisitos necesarios para optar al **Grado de Doctor**, a cuyos efectos será presentada en la Universidad Autónoma de Madrid, autorizando su presentación ante el Tribunal Calificador.

Y para que así conste se extiende el presente certificado,

Madrid, Enero 2017



V°B° de la Directora

Dra. Anna González Neira

V°B° del Tutor

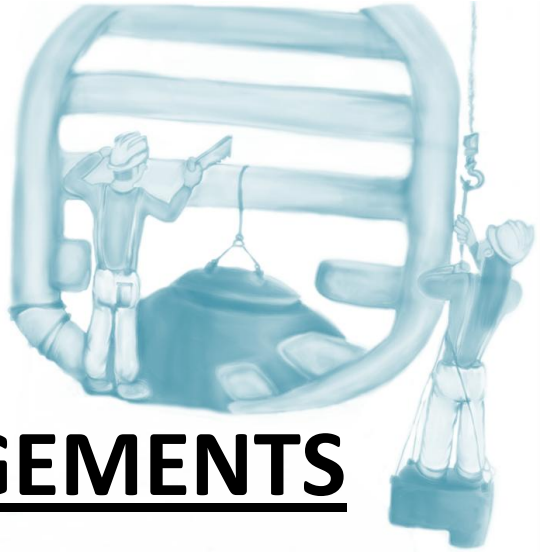
Dr. Sebastián Cerdán

This Doctoral Thesis has been elaborated in the Human Genotyping-CeGen Unit at the Spanish National Cancer Research Centre (CNIO) in Madrid between 2012 and 2017 under the supervision of Dr. Anna González Neira.

This work has been supported by the following grants and fellowships:

- Bancaja-Centre for Biomedical Network Research on Rare Diseases (CIBERER) 'Becas Lanzadera' ("Scholarship Shuttle") fellowship, 2010-2011
- Fundación Ramón Areces-Universidad Autónoma de Madrid M.Sc fellowship, 2010-2011
- Pediatric cancer research Project supported by Spanish Association Against Cancer (AECC), 2010-2013
- Project PI12/00226 supported by Institute of Health Carlos III (ISCIII), 2013-2016
- Severo Ochoa Excellence Programme (Project SEV-2011-191) PhD fellowship, 2014-2017

A mis padres, a mi hermana y a Dani



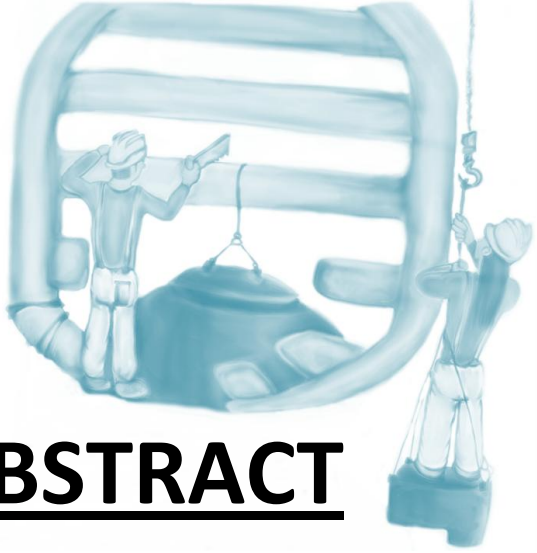
ACKNOWLEDGEMENTS

Quisiera agradecer a todas las personas que han participado de una u otra manera en la realización de este trabajo de tesis doctoral comenzando por Anna González y terminando por el resto de los miembros, pasados y presentes, de la Unidad de Genotipado Humano-CeGen y del programa de Genética del Cáncer Humano del Centro Nacional de Investigaciones Oncológicas.

¡¡Ha sido un placer!! ¡¡Muchas gracias por acompañarme en este viaje!!

En segundo lugar, pero no menos importante, muchas gracias a mis padres, a mi hermana y a Dani, a los que dedico este trabajo de tesis doctoral, porque lo más importante que tengo es vuestro afecto y vuestro apoyo incondicionales.

¡¡Muchas gracias!!



ABSTRACT

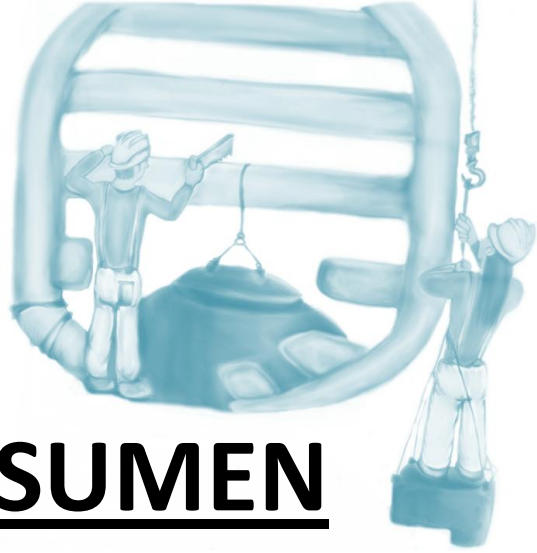
Patients vary widely in their response to medications. This is reflected in differences in treatment efficacy and toxicity and results in significant morbidity and mortality. The etiology of this observed variation is multifactorial, and can be explained in part by genetic factors. Gaining a better understanding of the relationship between human genetics and drug response is essential in cancer chemotherapy, as most chemotherapeutics have a narrow therapeutic window, and could lead to the development and adoption of personalized treatments. In this thesis we focused on the identification of germline genetic variants associated with treatment outcome in children diagnosed with Ewing sarcoma and risk of development of specific toxicities induced by the cytotoxic drug capecitabine (hand-foot syndrome) and by anthracyclines (chronic cardiotoxicity).

Despite the effectiveness of current treatment protocols for Ewing sarcoma, the prognosis for patients with metastatic or refractory disease is dismal. The identification of genomic biomarkers of response and survival is therefore of major importance in order to optimize treatment for these patients. We studied the genetic variation across the genes involved in the transport and metabolism of drugs used in Ewing sarcoma. We identified and replicated associations with overall survival for three common variants located in the *ABCC6*, *ABCB1* and *CYP2C8* genes.

Hand-foot syndrome is one of the most relevant dose-limiting adverse effects of capecitabine, experienced by more than 30% of patients and commonly leading to early discontinuation of capecitabine-based therapy. The few proven genetic markers of toxicity risk are variants involved in capecitabine biotransformation, but their predictive utility is uncertain. By combining genotyping-based genome-wide analysis and functional studies we identified a risk locus near the *CDH4* gene, encoding R-cadherin, strongly associated with severe capecitabine-induced hand-foot syndrome occurrence and prompting changes in *CDH4* gene expression, possibly through changes in chromatin topology. We also found that skin from patients who develop severe toxicity exhibited low levels of R-cadherin (highly expressed in the suprabasal granular layer of the epidermis) and involucrin (a component of the cornified envelope of the epidermis and essential for the skin barrier function) before capecitabine treatment.

Anthracyclines, which are widely used chemotherapeutic drugs, can cause progressive and irreversible cardiac damage and fatal heart failure in children and in adult cancer patients; established risk factors have been proven insufficient to accurately stratify patients. We found novel associations for low-frequency coding variants in the *GPR35* gene in pediatric oncology patients and in the *ETFB* and *WISP1* genes in anthracycline-treated patients, independently of age at diagnosis.

The studies performed as part of this thesis provide clear evidence that genetic variants could be used as predictors of drug efficacy and adverse drug reactions and could therefore be informative in the design of personalized cancer therapy.



RESUMEN

Es bien conocida la existencia de importantes diferencias interindividuales en la respuesta a la mayoría de los medicamentos, pero dicha variabilidad es de especial relevancia en oncología, ya que el cáncer es una de las principales causas de morbilidad y mortalidad y puesto que los antineoplásicos presentan un índice terapéutico muy estrecho. Aunque dicha variabilidad puede ser atribuida a diferentes factores, la variación genética se postula como uno de los factores más importantes. En esta tesis doctoral nos hemos centrado en la identificación de variantes genéticas asociadas a respuesta tumoral, supervivencia y efectos adversos en pacientes oncológicos.

A pesar de la eficacia de los tratamientos actuales para el sarcoma de Ewing, el pronóstico para aquellos pacientes con enfermedad metastásica y refractaria es desfavorable. Por tanto, la identificación de variantes genéticas predictivas y pronósticas resulta esencial para optimizar el tratamiento en dichos pacientes. Con este objetivo estudiamos variantes genéticas en aquellos genes implicados en el transporte y metabolismo de los fármacos más comúnmente utilizados en el tratamiento del sarcoma de Ewing e identificamos tres variantes comunes en los genes *ABCC6*, *ABCB1* y *CYP2C8* asociadas con la supervivencia de estos pacientes.

El síndrome mano-pie es el más común de los afectos adversos producidos por la capecitabina, afectando a más de un 30% de los pacientes. Su aparición conlleva graves consecuencias clínicas ya que obliga a reducir la dosis o incluso a la suspensión del tratamiento. Los marcadores genéticos identificados hasta la fecha se encuentran involucrados en su totalidad en el proceso de biotransformación de la capecitabina y su utilidad predictiva es incierta. La combinación de un estudio de asociación del genoma completo junto con estudios funcionales nos ha permitido la identificación de un locus de susceptibilidad a este efecto adverso. Este locus regula la expresión del gen *CDH4*, probablemente mediante alteraciones en la topología de la cromatina. Además encontramos que aquellos pacientes que sufren una toxicidad más severa presentan en su piel reducidos niveles de la proteína que codifica *CDH4,* R-cadherina (altamente expresada en el estrato granuloso de la epidermis) y de involucrina (proteína del estrato cornificado esencial para la función de barrera de la epidermis).

El tratamiento con antraciclinas produce efectos cardiotóxicos irreversibles e incluso fatales tanto en niños como en adultos y los factores de riesgo identificados hasta la fecha se han mostrado insuficientes para la estratificación de los pacientes. En esta tesis hemos encontrado nuevas variantes poco frecuentes asociadas a cardiotoxicidad en los genes *GPR35* en pacientes oncológicos infantiles, y *ETFB* y *WISP1* independientemente de la edad de diagnóstico.

Todos los resultados obtenidos en esta tesis doctoral suponen una clara evidencia del importante papel que juega la variación genética en las diferencias interindividuales observadas tanto en la eficacia como en la toxicidad de los quimioterapéuticos y por tanto, podrían ser utilizados en el futuro en el diseño de una medicina individualizada en estos pacientes.

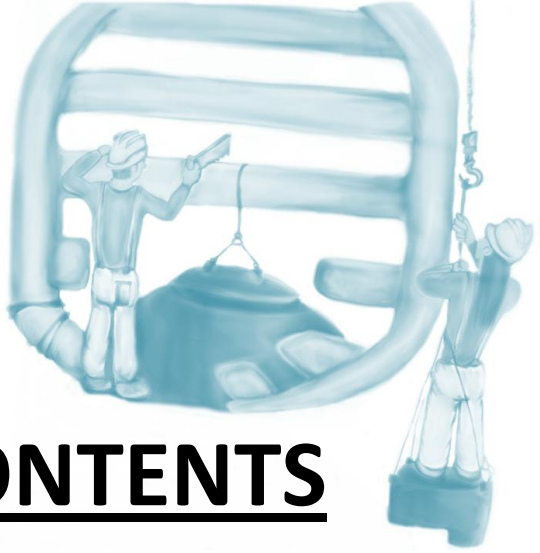
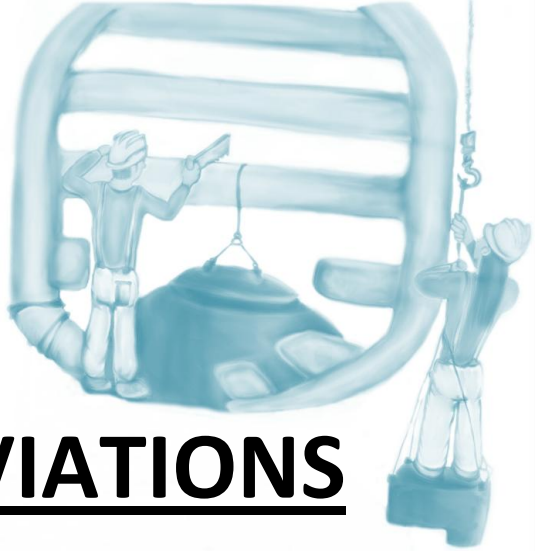


TABLE OF CONTENTS



ABBREVIATIONS

4C: circularized chromosome conformation capture

5′-DFCR: 5'-deoxy-5-fluorocytidine

5′-DFUR: 5'-deoxy-5-fluorouridine

5-FU: 5-fluorouracil

A: actinomycin-D

ABC: ATP-binding cassette

ABCB1: ATP-binding cassette sub-family B member 1

ABCC6*:* ATP-binding cassette sub-family C member 6

ADR: adverse drug reaction

AIC: anthracycline-induced cardiotoxicity

ASO: allele-specific oligonucleotide

bp: base pair

C: cyclophosphamide

CDA: cytidine deaminase

CDH4: cadherin 4

CDKN2A: cyclin dependent kinase inhibitor 2A

CES2: carboxylesterase 2

ChIA-PET: chromatin interaction analysis with paired-end tag

ChIP-seq: chromatin immunoprecipitation sequencing

chr: chromosome

CI: confidence interval

CiHFS: capecitabine-induced hand-foot syndrome

CNA: copy number alteration

CNV: copy number variation

CTCAE: common terminology criteria for adverse events

CTCF: CCCTC-binding factor

CYP2C8: cytochrome P450 family 2 subfamily C member 8

D: doxorubicin

DNA: deoxyribonucleic acid

DPYD: dihydropyrimidine deshydrogenase

E: etoposide

ECad: E-cadherin

ENCODE: Encyclopedia of DNA Elements

eQTL: expression quantitative trait locus

ETFB: electron transfer flavoprotein beta subunit

EWSR1: EWS RNA binding protein 1

FDR: false discovery rate

FDUMP: 5-fluoro-2′-deoxyuridine 5′-monophosphate

FLG: filaggrin

FUDR: 2′-deoxy-5-fluorouridine

FUTP: 5-fluorouridine-5′-triphosphate

GPR35: G protein-coupled receptor 35

GWAS: genome-wide association study

HLA: human leukocyte antigen

HR: hazard ratio

I: ifosfamide

IVL: involucrin

kb: kilobase

KD: knock-down

KRT1: keratin 1

KRT10: keratin 10

LD: linkage disequilibrium

LSO: locus specific oligonucleotide

LV: left ventricle

LVEF: left ventricle ejection fraction

MAF: minor allele frequency

mRNA: messenger RNA

MTHFR: methylenetetrahydrofolate reductase

NCI: National Cancer Institute

NHEK: normal human epidermal keratinocytes

OR: odds ratio

OS: overall survival

PCA: principal component analysis

PCR: polymerase chain reaction

QC: quality control

Q-Q plot: quantile-quantile plot

qRT-PCR: real-time quantitative polymerase chain reaction

RCad: R-cadherin

RNA: ribonucleic Acid

SF: shortening fraction

SKAT-O: optimized sequence kernel association test

SNV: single nucleotide variant

TAD: topological associating domain

tagSNV: tag single nucleotide variant

TK1: thymidine kinase 1

TMA: tissue microarray

TPMT: thiopurine S-methyltransferase

TYMP: thymidine phosphorylase

TYMS: thymidylate synthetase

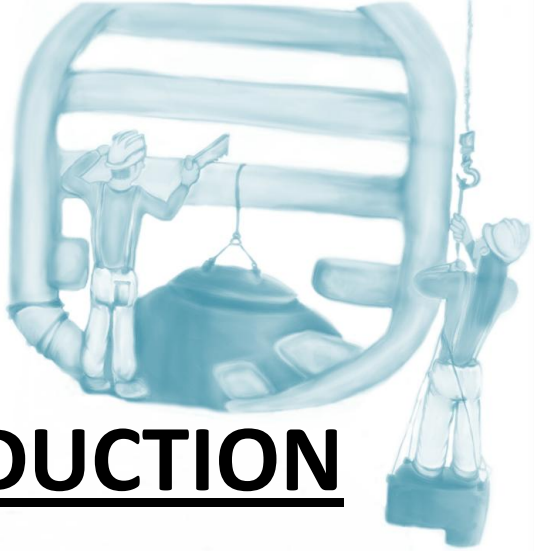
UPP1: uridine phosphorylase 1

V: vincristine

***VEGF*:** vascular endothelial growth factor

WES: whole-exome sequencing

WISP1: WNT1 inducible signaling pathway protein 1



INTRODUCTION

1. Identification of genomic biomarkers

There is great heterogeneity in the way individuals respond to clinical drugs in terms of both, treatment efficacy and host toxicity. A drug that is effective in one person may have no discernible therapeutic effect in another, and in some may result in undesirable or fatal side-effects, even though the medication is administered at a normal recommended dose[1]. Both drug inefficacy and adverse drug reactions (ADRs) remain a major clinical problem. It has been estimated that the non-response rate to a major drug among patients diagnosed with several important diseases ranges from 20% to 75%, with the lowest rate for cancer chemotherapy[2]. On the other hand, ADRs are one of most common causes of drug withdrawal, accounting for 7% of all hospital admissions and responsible for >100,000 deaths per year, making ADRs between the fourth and sixth leading cause of death in developed countries[3,4]. This is particularly important in oncology because cancer is one of the leading causes of morbidity and mortality worldwide; and chemotherapeutic agents, which in general affect tumor and non-tumor cells, have a narrow therapeutic index, with the potential for life-threatening toxicity and where treatment discontinuation is often fatal[5]. Thus, there is an urgent need for novel treatment strategies that can improve cure rates and decrease adverse events in oncology.

Although there are multiple contributory factors influencing the effect of drugs, such as physiopatological factors (e.g., age, gender, body mass index, organ function, concomitant diseases) and environmental variables (e.g., nutritional factors, alcohol consumption), genetic factors can account for 20% to 40% of interindividual heterogeneity in drug efficacy and play a significant role in the incidence and severity of ADRs[6]. In fact, for certain drugs or drug classes, genetic factors have been shown to be the most important influence on drug treatment outcome[7]. Over the last decades rapid advances in research have greatly increased our understanding of the molecular basis and genetics of tumor progression and drug response, and these advances have also led to the identification of numerous genomic biomarkers in oncology. In general, a biomarker is defined as a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention[8]. According to U.S Food and Drug Administration (FDA) and European Medicines Agency (EMA) a genomic biomarker is defined as a measurable DNA and/or RNA characteristic that is an indicator of normal biologic or pathogenic processes and/or response to therapeutic or other interventions[9,10]. Potential genomic biomarkers could, for example, be in genes involved in the pharmacokinetics (drug absorption, distribution,

metabolism and elimination) or pharmacodynamics (effects on drug receptors and other drug targets) of medications. DNA characteristics include, but are not limited to: single nucleotide variants (SNVs), variability of short sequence repeats, DNA modifications (e.g., methylation), deletions or insertions, copy number variations (CNVs) and cytogenetic rearrangements (e.g., translocations, duplications, deletions or inversions). RNA characteristics include, but are not limited to: RNA sequences, expression levels or processing (e.g., splicing and editing) and microRNA levels.

In general, biomarkers can be divided into two types: prognostic and predictive. Prognostic markers aim to objectively predict the patient's clinical outcome, such as the probability of cancer recurrence after standard treatment. Predictive biomarkers aim to objectively predict the response of a patient to a specific clinical intervention and are associated with tumor sensitivity or resistance to that therapy, including toxicities[11,12]. Cancer treatment is especially complex because a combination of inherited variations within the individual (germline) and acquired variations within the tumor (somatic) influence disease outcome and the response/ toxicity to the drug therapy[13,14]. Germline variation may alter drug pharmacokinetics and pharmacodynamics leading to toxicity and/or lack of efficacy. On the other hand, somatic alterations, but also germline changes can be relevant to predict tumor response [15,16]

In this thesis we focused on the identification of predictive and prognostic germline SNVs for cancer treatment. SNVs are single-nucleotide substitutions of one nucleotide for another and are by far the most common genetic alteration. SNVs can occur in both coding and non-coding regions of the genome, and when located in a coding region, they can alter the amino acid sequence (non-synonymous or missense variants) or be silent (synonymous variants). SNVs may influence, for example, gene expression, messenger RNA (mRNA) stability, gene splicing, transcription factor binding, subcellular localization of mRNAs and/or proteins, or the sequence of non-coding RNA[17].

The ability to identify genomic biomarkers corresponding to a therapeutic effect is the basis for the concept of the so-called personalized medicine. The National Cancer Institute (NCI) has defined personalized medicine …"as a form of medicine that uses information about a person's genes, proteins and environment to prevent, diagnose, and treat disease"[18]. Rather than having a unique treatment for each individual person, the ultimate goal of personalized medicine is to identify patients who are drug responders and patients who are prone to develop drug toxicity;

that is, effective and safe medication to targeted patients with appropriate genotypes[19,20]. Achieving individualized medicine requires a deep understanding of the genomic and molecular basis underlying variability in drug response, especially in complex diseases, such as cancer. Large international projects such as the 1000 Genomes Project[21], the Cancer Genome Atlas[22], or the International Cancer Genome Project[23], are producing increasingly comprehensive maps outlining the regions of the human and cancer genomes containing variants. These international initiatives along with the high-throughput "omics" technologies´ evolution will help us move from population-based prescribing toward individualized cancer treatment; with the ultimate goal of improving treatment efficacy, reduce toxicity and minimize cost.

2. Strategies to identify genomic biomarkers

In general there are two approaches used to evaluate how genetic variation contributes to interindividual variability in drug response (drug efficacy and related-toxicity): candidate gene and genome-wide.

2.1. Candidate gene analysis

Candidate gene approaches focus on one or more candidate genes or pathways and are hypothesis-driven. Selection of putative genes in the candidate gene approach is based in the current knowledge of human pathophysiology, pharmacology, and cancer biology. In general, genetic variants are selected for having a functional consequence in our candidate gene, either by affecting gene regulation or its protein product is selected. The hypothesis is that the genetic variants, mostly within a gene with a relevant role in the drug pharmacokinetics and pharmacodynamics, would affect the drug´s efficacy and/or toxicity and these studies are therefore also termed hypothesis-driven association studies. Typical candidate genes encode, for example, drug transporters, biotransformation enzymes, or drug receptors; and clinically relevant samples (e.g., blood, tissues or tumor specimens), which represent either drug toxicity or functional sites, are used. A classical example of a clinically relevant candidate gene is *TPMT* (thiopurine S-methyltransferase) in which three non-synonymous common variants*, TPMT*2*, *TPMT*3A*, and *TMPT*3C,* account for 80%–95% of the lower TPMT enzyme activity[24]. The positive finding of a candidate gene approach is easy to interpret and can yield clinically relevant information, but the candidate gene strategy carries the risk of not finding an association, likely due to small simple size of cohorts used to detect an effect; and the exclusion of important or causal variants or genes[25–27].

However, in most of the cases, one single gene/genetic variant is not sufficient to explain the wide interindividual differences observed in drug efficacy or toxicity; which are often polygenic, with several genes within and across many pathways involved. For this reason, it seems obvious the advantage of conducting a candidate pathway approach; combining information on several genes that are common to a pathway; although there is still a risk of excluding other relevant genes. The pharmacological pathways of methotrexate and mercaptopurine illustrate the potential of this kind of approach[28].

Besides genotyping-based candidate gene approaches, next-generation sequencing is currently being used to perform candidate-gene strategies by targeted enrichment of a set of genes and exclusively performing sequencing on them (gene panels). Gene panels represent a cost-effective alternative to whole-genome and whole-exome sequencing (WES) and entail considerably lower sequencing costs, but will only succeed if the trait-causing gene is included in the panel. Furthermore, targeted-gene sequencing is at the forefront of the current efforts given in the application of routine diagnostics into the clinic[29–31].

In this thesis we adopted a genotyping-based candidate pathway strategy to identify genetic variants that predict treatment outcome in children diagnosed with Ewing sarcoma (**Study I**).

2.2. Genome-wide association analysis

In contrast to candidate gene strategies, a genome-wide approach gives equal weight to all genes in the genome, is hypothesis-generating, and can be used when little is known regarding gene-drug effect; helping to prioritize genes or genomic regions for further investigation[27].

Genome-wide association studies (GWAS) using genotyping arrays have evolved over the last ten years into a powerful tool for investigating the genetic architecture of complex diseases, including drug response to chemotherapeutics. These studies analyze hundreds of thousands of common SNVs across the entire human genome. Therefore, these studies do not make use of the current knowledge about gene function or a drug´s mechanism of action and they are hypothesis-generating[32,33].

Genotyping arrays used in GWAS studies include mostly tagSNVs, which are representative SNVs sufficient to capture most of the haplotype structure of a particular region in high linkage

disequilibrium with them. Due to the very large number of SNVs interrogated in a GWAS, they require much larger sample sizes to achieve an adequate statistical power[34] and positive findings always need to be replicated in independent series of patients[35] in order to avoid false positive findings. Given that most of GWAS tagSNVs (approximately 88%)[36] are located in intergenic or intronic regions a major challenge lies in the recognition of the causal variant responsible for the observed association; as well as the interpretation of study results, requiring fine mapping and mechanistic studies to understand the biological plausibility of certain findings[32]. Biological interpretation of GWAS findings is even more difficult in cancer pharmacogenetics (the study of the genetic factors that influence drug efficacy and toxicity[37]), since drug response represent a polygenic phenotype and in most cases they are measured in the form of a subjective, ordinal scale (grade); and the complexity and the multifactorial nature of cancer with many intrinsic and extrinsic factors involved. A useful approach to overcome this problem is to select patients with extremely differentiated phenotypes at both ends of the phenotype, which have a greater chance of carrying the causal genetic variant/s contributing to the trait, rather than to include large series of patients that might dilute gene-phenotype associations[38–40].

GWAS have been extensively applied in pharmacogenetics studies and they have helped in the identification of novel associations that would be unlikely to have been detected by candidate gene studies. The primary area in which this methodology was used in pharmacogenetics was in drug efficacy (approximately 70% of published GWAS in pharmacogenetics), whereas the remaining studies focused on ADRs[32,36]. As an example, using a genome-wide interrogation of germline variation, Treviño et al[41]. uncovered genetic variants that affected the disposition and effects of methotrexate in children with acute lymphoblastic leukemia in *SLCO1B1*, a gene not previously investigated as a candidate gene in clinical pharmacogenetic studies of methotrexate.

In the current thesis we performed a genotyping-based genome-wide study applying extreme phenotype sampling to identify capecitabine-induced hand-foot syndrome (CiHFS) susceptibility variants (**Study II**).

Genotyping-based GWAS have greatly improved our understanding of the genetic basis of multitude of complex traits and diseases and they have identified more than 29,000 disease-associated variants[36]. However, most of these associated SNVs are common variants having a population frequency of 5% or greater, and they usually explain only a small proportion of the

genetic variance of the trait[42,43]; leading to the possibility that low-frequency and rare variants could explain additional risk or trait variability and play an important role in uncovering the so-called "missing heritability". Low-frequency and rare genetic variants are likely to have arisen from mutation events and, *a priori,* they are expected to have larger effects on complex traits than common variants because they will not have been subject to negative selection[44]. The most comprehensive approach to characterize low-frequency and rare genetic variation is through whole-genome sequencing in large number of individuals. However, the combination of large scale whole-genome sequencing and classical association studies is impractical for many research groups because of the high cost. Until whole-genome sequencing is inexpensive enough to be generally used in large samples, various approaches have arisen as principal alternatives for sequencing-based genome-wide studies, such as WES, and whole-exome genotyping arrays[45]. Both, WES and whole-exome genotyping arrays focused on the protein-coding regions of the genome, which represent a very small portion of the whole genome (around 1%) but are more likely to affect protein function and thereby have an impact on human traits and disease[46]. In the current thesis we performed an exome-array genome-wide analysis to identify genetic variants predictive of susceptibility to chronic anthracycline-induced cardiotoxicity (AIC) in pediatric and breast cancer patients (**Study III** and **Study IV**)

The exome content of whole-exome genotyping arrays, such as the Illumina Infinium HumanExome BeadChip used in this thesis, have been selected from over >12,000 individual exome and whole-genome sequences representing diverse ethnicities and a range of common traits (cancer, type 2 diabetes, metabolic and psychiatric disorders). Apart from non-synonymous and splicing and stop-altering variants, exonic genotyping arrays contain additional variants, including variants associated with complex traits in previous GWAS, ancestry-informative markers, human leukocyte antigen (HLA) tagSNVs and mitochondrial SNVs[47]. Although exome arrays constitute a cost-effective alternative to genome-wide sequencing in large number of individuals, substantially increasing statistical power for variants that are on the chip; the limitation of these chips compared with whole-exome sequencing is that they does not provide complete coverage of all coding variants at each locus and of ethnic-specific variants, since individuals with European ancestry are the predominant source of variation in these arrays. However, the utility of these arrays has been clearly demonstrated and during the last three years the first results of exome chip-based studies have been published, and novel associations for low-frequency and rare coding variants with complex traits such as insulin processing[48], asthma[49], liver disease[50], type 2 diabetes[51], or schizophrenia[52] have been found.

2.3. Novel statistical association analyses

Associations between genetic variants and a particular trait are typically evaluated by linear regression for continuous traits and by logistic regression for binary traits; however, single-variants test is adequate for common variants but is often underpowered to detect rare variant associations. The development of appropriate statistical methods to analyze rare variant associations to achieve an adequate statistical power has become a very active research area[45,53]. Methods for rare variant association testing have mainly focused on the analyses of rare variants within the same functional region (e.g., gene or pathway) and then consider their joint effects on complex traits. Because rare variants are traditionally grouped by genes, these tests are referred to as gene-based tests. Gene-based tests can be divided into two broad categories: burden and variance-component tests. Burden tests collapse or summarize the rare variants within a region into a single genetic variable, which can then be tested for association with the trait. The main limitation of burden tests is they assume that all variants within the collapsed regions are causal and influence the trait in the same direction and magnitude of effect[45]. However, in practice variants with protective and deleterious effects exist and the magnitude of each variant´s effect is likely to vary. To address these limitations, variance-component tests such as SKAT[54], were developed. Variance component tests avoid the directionality of effect and consequently are more powerful than burden tests if a region has many noncausal variants or if the variants have different directions of association[53,54]. However, if a large proportion of the rare variants in a region are truly causal with the same direction of association, then burden tests are more powerful. Because both scenarios can arise and the true underlying genetic model is in most cases unknown, combined tests unifying burden and variance-component tests, such as the SKAT-O[55], are desirable.

In the current thesis we used gene-based tests (SKAT-O) to identify new genes and novel low-frequency and rare genetic variants influencing the susceptibility to chronic anthracycline-induced cardiotoxicity (AIC) (**Study III** and **Study IV**).

3. Identification of predictive and prognostic genomic biomarkers for Ewing sarcoma (Study I)

In Study I we focused on the identification of predictive and prognostic germline genetic variants for Ewing sarcoma.

Ewing sarcoma is a rare, highly cellular malignant round-cell tumor of bone and soft tissue primarily affecting children and adolescents, with a peak incidence at age 15. It is also called the Ewing sarcoma family of tumors and includes Ewing sarcoma of bone, extraosseous Ewing sarcoma, peripheral primitive neuroectodermal and Askin's tumors[56]. Despite being the second most frequent primary malignant bone tumor after osteosarcoma, Ewing sarcoma accounts for only 2.9% of all childhood cancers, with an annual incidence of 2.5-3 cases per million[57].

The majority of Ewing sarcoma tumors develop in the diaphyseal regions of long bones, such as the femur; followed by the pelvis, chest wall and the spine; although, approximately 20% arise in soft tissues. Ewing sarcoma is an aggressive tumor with a high incidence of metastasis at presentation ranging from 20%-25% (10 % lung, 10 % bones/bone marrow, 5 % combinations or others)[58]. This pediatric tumor occurs predominantly in Caucasians, with very few cases in African and Asian populations[59] and is slightly more common in males than in females (ratio 55:45). Interestingly, it has been found significant racial and ethnic differences in the age, primary tumor site (bone *v* soft tissue), and tumor size in Ewing sarcoma patients suggesting the presence of genetic factors for Ewing sarcoma susceptibility[59,60].

Current therapy for Ewing sarcoma consists of initial chemotherapy after biopsy (neoadjuvant chemotherapy) to eradicate systemic disease; then local control with surgery and/or radiation, followed by chemotherapy (adjuvant or consolidation chemotherapy) to treat, not only the primary tumor, but also presumed microscopic metastases and, to prevent recurrence. Chemotherapy for Ewing sarcoma consists of a 6-drug backbone of vincristine, doxorubicin and cyclophosphamide alternating with ifosfamide, etoposide and actinomycin-D[58]. With the use of these modern multimodal regimens, cure rates of around 70% can be achieved in patients with localized disease, however, they are ranging between 20%-40% for those with metastatic Ewing sarcoma. Furthermore, 30%-40% of patients with a localized primary tumor and 60%-80% of patients with disseminated disease experience relapse (either locally or distantly, or both) after treatment and have a dismal prognosis, with a likelihood of long-term survival after recurrence lower than 15%[61,62]. Thus, the main challenge for Ewing sarcoma remains in preventing recurrence and drug resistance and improving outcome, especially in those patients with metastatic and relapsed/recurrent disease.

The majority of biomarkers studied in Ewing sarcoma are prognostic and nearly all are somatic biomarkers[63]. The outcome of Ewing sarcoma tumors is influenced by many clinical-

pathologic and treatment-related factors. Central primary tumor location (axial *v* appendicular), increasing tumor size and patient age at diagnosis and decreased tumor necrosis after neoadjuvant chemotherapy have all been implicated as negative prognostic features; however the presence of metastatic disease at diagnosis is currently the strongest prognostic biomarker in Ewing sarcoma and has been proven in all clinical studies and routine clinical practices[58,61,64]. Outside of above mentioned prognostic factors, many efforts have been made to identify potential biomarkers in Ewing sarcoma. Ewing sarcoma is specifically characterized by a translocation that fuses the *EWSR1* (EWS RNA binding protein 1) gene to one of the several different genes belonging to the ETS family, being the *EWSR1-FLI1* translocation (also known type 1 fusion) the most common fusion type observed (~85% of Ewing sarcoma cases)[65]. Early retrospective studies reported better outcome for tumors harboring a type 1 fusion, compared with those with other fusion types[66,67]; however, prospective studies evaluating *EWSR1* fusion status and patient outcome failed to confirm the original findings[68].

On the other hand, genetic alterations affecting cell-cycle proteins, including *CDKN2A* (cyclin dependent kinase inhibitor 2A) (*INK4A/ARF* locus) deletion and *TP53* (tumor protein p53) overexpression or mutations [especially when *STAG2* (stromal antigen 2) and TP53 co-ocurred] have been associated with poor prognosis in Ewing sarcoma patients[69,70]. On the other hand, several of the most frequent copy number alterations (CNAs) observed in Ewing sarcoma tumors, such as the gain of the whole chromosome 8 and chromosome 12, the gain of the long arm of chromosome 1 and deletion of the long arm of chromosome 16[69] have been also correlated with poor prognosis[58,64], but a prospective analysis of CNAs and clinical outcome has not yet been undertaken.

Recent studies have used flow cytometry to identify circulating tumor cells in peripheral blood and bone marrow samples from Ewing sarcoma patients and these studies provide preliminary support for the potential prognostic significance of circulating tumor cells, especially for subclinical disease[71,72]. Many other prognostic markers in Ewing sarcoma have been studied and associated with significant differences in outcome, including biomarkers related with angiogenes, such as VEGF (vascular endothelial growth factor)[73]; or tumor microenvironment, such as the chemokine receptor CXCR4[74]; however most of these studies were retrospective and/or have been conducted in small cohorts of patients[58,64].

Regarding predictive biomarkers, the expression of the cell surface protein CD133 has been associated with chemoresistant Ewing sarcoma disease[75] and the expression and nuclear localization of the insulin receptor IGF1R has been identified as a putative marker of treatment response[76].

Given the lack of candidate gene studies in Ewing sarcoma exploring the role of germline variants on treatment outcome (on both, tumor response and overall survival), in this thesis we focused on the identification of genetic variants in genes involved in the pharmacokinetics of the chemotherapeutic drugs most commonly used to treat Ewing sarcoma.

4. Predictive genomic biomarkers for ADRs

4.1. Identification of genetic variants predictive of susceptibility to capecitabine-induced hand-foot syndrome (CiHFS) (Study II)

Capecitabine is and oral prodrug of 5-fluorouracil (5-FU), which is currently approved for its use as a single agent for patients with metastatic breast cancer that is resistant to both paclitaxel and anthracyclines, and for those with breast cancer resistant to paclitaxel and for whom further anthracycline therapy is contraindicated. Capecitabine is also approved for combination therapy with docetaxel for the treatment of patients with metastatic or locally advanced breast cancer in whom prior anthracycline-based therapy has failed. It is also approved as monotherapy in the adjuvant setting when treating Dukes' stage C colon cancer and patients with metastatic colorectal cancer, as well as in the adjuvant setting in combination with platinum-based therapy for patients with advanced or metastatic colorectal tumors[77].

Capecitabine has a number of advantages over traditional 5-FU: i) it is orally administered, enabling dosing that approximates to continuous infusion 5-FU but in a more convenient outpatient setting, avoiding the inconvenience and complications associated with central venous catheters or infusion pumps[78,79]; ii) it has increased tumor selectivity and thus, reduced toxicity; and iii) improved efficacy[80].

After absorption across the digestive tract, capecitabine is metabolized and activated to 5-FU through three sequential enzymatic reactions (**Figure I1**). In the first step, capecitabine is hydrolysed to 5'-deoxy-5-fluorocytidine (5'-DFCR) by carboxylesterase (CES) in the liver, which is

further converted to 5′-deoxy-5-fluorouridine (5′-DFUR) via cytidine deaminase (CDA) a ubiquitous enzyme which is found in high concentrations in liver, intestine, plasma and tumor tissue. As part of the capecitabine´s rationally designed activation, 5′-DFUR conversion to 5-FU preferentially occurs in tumor tissue via thymidine phosphorylase (TYMP)[81]. It has been shown that TYMP is present in higher concentrations in tumor tissue (2.5-fold more) compared with normal adjacent tissues, and 14-fold more than in the plasma[82], thus theoretically providing tumor specificity and limiting systemic toxicity. 5-FU is further activated in the tumor to cytotoxic active metabolites that inhibit DNA and protein synthesis through the binding to thymidylate synthetase (TYMS) and the incorporation of metabolites as false nucleotides into DNA and RNA, leading to cell proliferation inhibition. 5-FU can be converted to its active metabolites 5-fluoro-2′-deoxyuridine 5′-monophosphate (FDUMP) and 5-fluorouridine-5′-triphosphate (FUTP) in both normal and tumor cells but it occurs at higher rates in rapid proliferating cells, such as malignant cells[82,83]. Finally, 5-FU is catabolised in the liver into dihydro-5-fluorouracil by the dihydropyrimidine deshydrogenase (DPYD) enzyme[82].

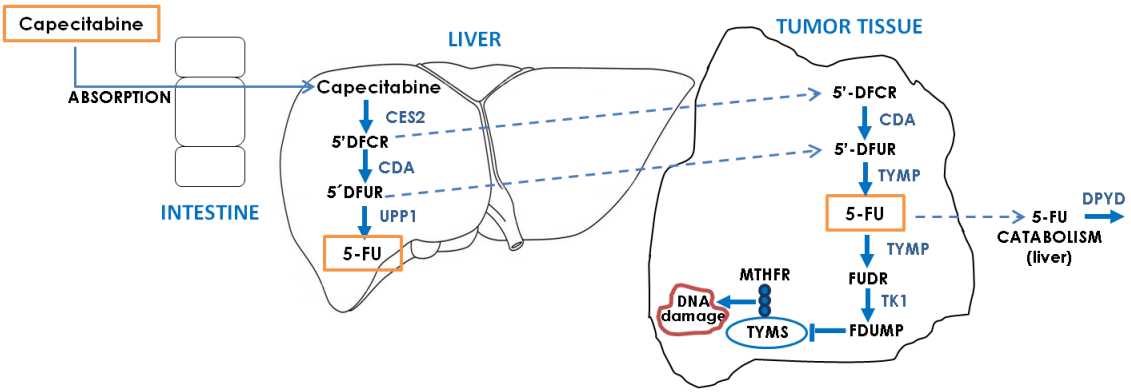


Figure I1. Enzymatic activation of capecitabine. Following oral administration, capecitabine is metabolized and activated to 5-FU via a three-step enzymatic process. Capecitabine was rationally designed so that concentrations of the cytotoxic metabolites FDUMP (5-fluoro-2′-deoxyuridine 5′-monophosphate) and FUTP (5-fluorouridine-5′-triphosphate) are higher within malignant cells than within normal cells. Abbreviations: 5-FU, 5-fluorouracil; CES2, carboxylesterase 2; 5′-DFCR, 5'-deoxy-5-fluorocytidine; CDA, cytidine deaminase; 5′-DFUR, 5'-deoxy-5-fluorouridine; UPP1, uridine phosphorylase 1; TYMP, thymidine phosphorylase; FUDR, 2′-deoxy-5-fluorouridine; TK1, thymidine kinase 1; TYMS, thymidylate synthetase; MTHFR, methylenetetrahydrofolate reductase; DPYD, dihydropyrimidine deshydrogenase.

Capecitabine is a generally well-tolerated cytotoxic and has an improved tolerability profile compared with bolus 5-FU/leucovorin[80], however its clinical use is limited by the appearance of adverse events, being the most common dose-limiting toxicities diarrhea and hand–foot syndrome. Other frequent adverse events include hyperbilirubinaemia, fatigue/weakness, and other gastrointestinal effects such as nausea/vomiting, abdominal pain and stomatitis/mucositis.

The incidence of adverse events in breast and colorectal cancer patients treated with capecitabine is very similar as well as the safety profile for capecitabine administered as monotherapy or in combination[80,84,85].

Table I1. Capecitabine-induced hand–foot syndrome (CiHFS) grading scale

Grade	Signs and symptoms	Presentation
Grade 1	Numbness, dysaesthesia, paresthesia, tingling, painless swelling, erythema or discomfort of the hands and/or the feet that does not interrupt normal activity	
Grade 2	Painful erythema and swelling of the hands or the feet, resulting in interruption to daily routine	
Grade 3	Desquamation, severe blistering and ulceration causing severe pain and discomfort of the hands and/or the feet, resulting in the inability to perform activities of day living	 

Capecitabine-induced hand–foot syndrome (CiHFS), also called palmar-plantar erythrodysaesthesia, is a cutaneous adverse event and the most frequent cause for dose reduction or therapy discontinuation. CiHFS is characterized by tenderness, redness and swelling and, in more severe cases, blistering, ulceration, desquamation or severe pain on the palms of the hands and/or the soles of the feet. CiHFS can appear in three different grades of severity

according to the NCI Common Terminology Criteria for Adverse Events (CTAE)[86] (**Table I1**), and hands are usually more affected than the feet. Grade 1 toxicity requires

close monitoring, but no therapy interruption. For grade 2 and grade 3 toxicities capecitabine administration should be stopped until resolution of symptoms or diminution to grade 1. For prolonged appearance of CiHFS and grade 3 toxicity dose reductions for subsequent cycles are recommended and even treatment withdrawal.

Treatment interruption and dose reduction usually lead to rapid reversal of symptoms without long-term consequences and CiHFS is never life-threating[80,84,85,87]. The overall incidence of CiHFS is around 30%, although it has been reported as high as 50% on clinical trials; with 17% of patients reporting a severe grade 3 form[80,85]. The severity of CiHFS is dependent on age, gender, local clinical practice and, possibly diet[88–90]; but also is influenced by genetic factors. Over the last decade a number of publications have identified common polymorphisms, but also rare variants contributing to the susceptibility to CiHFS[91–99]. However, all of these studies have focused on genes involved in the biochemical pathway of capecitabine activation and subsequent 5-FU action and catabolism, and the majority throughout a candidate gene approach. Although these studies provided important biological and clinical information, they do not fully explain the whole spectrum of the interindividual variabilities observed in CiHFS[97]. In addition, the biological mechanisms underlying CiHFS remain poorly understood. Thus, in this thesis we focused on the identification of additional genetic susceptibility variants by using a genotyping-based genome-wide approach coupled with a selection of extreme phenotype breast and colorectal capecitabine-treated patients, in order to ensure accurate patient phenotyping and to avoid the subjectivity associated to grading scales of CiHFS.

4.2. Identification of genetic variants predictive of susceptibility to chronic anthracycline-induced cardiotoxicity (AIC) (Study III and Study IV)

Daunorubicin was the first anthracycline antibiotic to be characterized and was isolated from *Streptomyces peucetius* in 1963[100]. Daunorubicin was found to be quite effective in treating leukemias and lymphomas[101]. Doxorubicin (also known as adriamycin) was identified a few years later, and shown to be a more effective anticancer drug[102]. After that, many other anthracyclines have subsequently become available, such as epirubicin or idarubicin. Anthracyclines are currently among the most widely used chemotherapeutics agents in children and adults and they are used in a broad variety of hematological malignancies, solid tumors (e.g., breast, lung, ovarian cancers), soft-tissue and bone sarcomas[103]. Despite their efficacy and widespread use,

anthracyclines are among the most notorious anti-neoplastic agents that cause cardiotoxicity[104]. Along with cancer recurrence and secondary malignancies, cardiovascular disease is the leading cause of morbidity and mortality among long-term cancer survivors[105,106]. Children cancer survivors are at higher risk than adults for anthracycline-induced cardiotoxicity (AIC), given that anthracyclines are more extensively used in children and young adolescents (more than half of the children with cancer receive an anthracycline as part of their treatment therapy[107]) and their developing cardiovascular system is particularly vulnerable to cardiotoxic effects of anthracyclines[108].

The mechanisms underlying AIC remain controversial and poorly understood, although they are probably multifactorial. Oxidative stress is the most study and widely accepted cause of AIC, although mitochondrial damage, impairment of mitochondrial calcium homeostasis and inhibition of sarcoplasmic reticulum function have all been also implicated in AIC[109–111]. The heart may be particularly vulnerable to anthracycline-induced damage for several reasons. Cardiomyocytes have an abundance of mitochondria and high concentrations of cardiolipin, a phospholipid found in abundance on the inner mitochondrial membrane and with high affinity for anthracyclines; allowing anthracyclines to enter mitochondria[112]. In addition, cardiomyocytes have low levels of key antioxidant enzymes, such as catalase and glutathione peroxidase, and therefore are especially vulnerable to oxidative stress. Anthracyclines have also been found to inhibit some of these enzymes, such as cardiac glutathione peroxidase, making cardiomyocytes even more susceptible to anthracycline damage[113].

AIC can be categorized as acute or chronic cardiotoxicity based on the time that symptoms manifest (**Table I2**)[103,104]. Acute cardiotoxicity usually occurs within 1 week of therapy and it is characterized by depression of contractile function. With current anthracycline protocols, acute cardiotoxicity is a rare event, occurring in less than 1%[114] of patients, and symptoms usually resolve with anthracycline therapy discontinuation. Chronic toxicity can be divided in two forms depending on the time of onset: early- (occurring the first year after anthracycline administration) and late-onset cardiotoxicity (occurring more than 1 year after anthracycline treatment). Both chronic forms are characterized by an irreversible left ventricle (LV) dysfunction that can be progressive and, which, in some cases leads to heart failure. Although chronic AIC takes the form of dilated cardiomyopathy in adults, children tend to present initially a dilated cardiomyopathy y that progressively becomes a restrictive cardiomyopathy[115] (**Figure I2**).

Table I2. Characteristics and course of the different types of AIC

Characteristic	Acute cardiotoxicity	Early-onset, chronic progressive cardiotoxicity	Late-onset, chronic progressive cardiotoxicity
Onset	Within the first week of anthracycline treatment	<1 year after the completion of anthracycline treatment	≥1 year after the completion of anthracycline treatment
Clinical features in adults	Transient depression of myocardial contractility	Dilated cardiomyopathy	Dilated cardiomyopathy
Clinical features in children	Transient depression of myocardial contractility	Restrictive cardiomyopathy, dilated cardiomyopathy, or both	Restrictive cardiomyopathy, dilated cardiomyopathy, or both
Course	Usually reversible when anthracycline treatment is discontinued	Can be progressive. Irreversible	Can be progressive. Irreversible

Early-onset AIC occurs in 1.6-2.1% of anthracycline-treated patients[116]. Remarkably, late-onset AIC may not become apparent until to 10 to 20 years after the first anthracycline dose and risk of cardiac events can persist up to 45 years after therapy completion[105,117]. In fact, the 30-year incidence of severe cardiac events for childhood cancer survivors treated with anthracycline is 7.3%[111]. The Childhood Cancer Survivor Study, a study enrolling 14,358 5-year survivors of childhood malignancies, revealed that the use of <250 mg/m^2 of anthracycline was associated with a 2.4-fold higher risk of developing congestive heart failure compared to those non anthracycline-treated patients 30 years after cancer diagnosis. When patients were treated with ≥250 mg/m^2, this risk increased to 5.2-fold, and 1.8-fold and 2.3-fold for pericardial disease and valvular abnormalities, respectively[110]. Long-term cardiotoxicity has not been extensively studied in long-term survivors of adult malignancies, but the French Adjuvant Study Group reported that among 3,577 early breast cancer survivors, the 7-year risk of LV dysfunction was 1.36% in epirubicin-treated patients compared to 0.21% in non-epirubicin-treated patients[118]. Independently if they are children or adults, survivors with late-onset AIC can be asymptomatic for several years before experiencing a symptomatic cardiac dysfunction, with subclinical changes in LV structure and function only detectable through echocardiography (up to 57% of

patients)[119]. Thus, the main challenge for anthracycline therapy remains in preventing subclinical damage and chronic symptomatic cardiotoxicity.

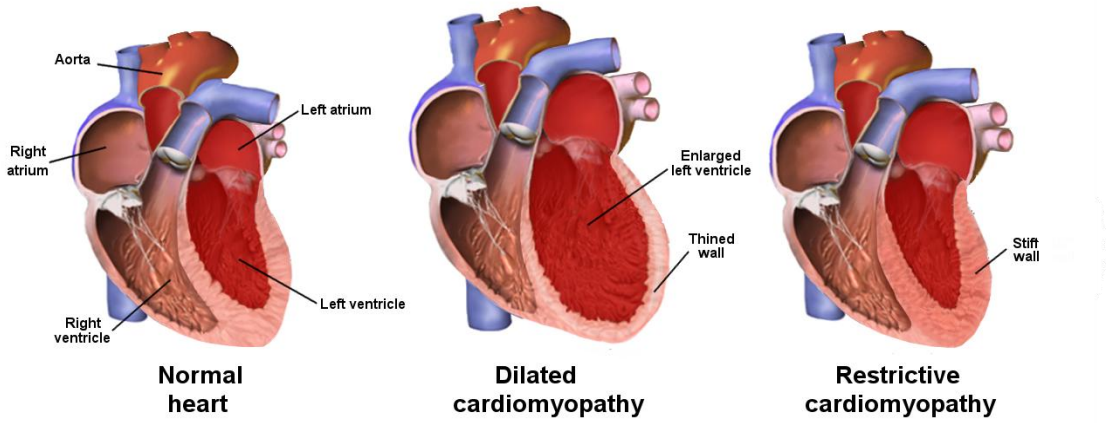


Figure I2. Characteristics of dilated and restrictive cardiomyopathies. In dilated cardiomyopathy, the left ventricle is enlarged (in some cases the right ventricle will also be larger) and has thin walls. Thus, the heart does not contract well; causing systolic dysfunction. In restrictive cardiomyopathy, the left ventricle is slightly smaller in chamber size and has rigid walls. As a result, the LV relaxes abnormally and does not properly fill with blood, leading to diastolic dysfunction. Figure adapted from Blausen gallery 2014 (DOI:10.15347/wjm/2014.010).

On the other hand, echocardiography is the most commonly used technique to monitor cardiac structure and function after anthracycline administration, and LV ejection fraction (LVEF) and shortening fraction (SF) are the most common echocardiographic indices measured[120]; however, in most cases echocardiography lacks the sensitivity and specificity to detect symptomatic cardiotoxicity and especially subclinical cardiac dysfunction[108,120,121]. As an alternative, serum biomarkers such as B-type natriuretic peptide and cardiac troponins (T and I) are increasingly being used to evaluate AIC during and after treatment[120,121]. Regarding primary prevention, there are two main strategies for prevention of AIC: i) reduce cardiotoxic potency by administering via continuous infusion, liposome encapsulation, or using a less cardiotoxic drug (e.g., epirubicin or idarubicin) and ii) use a cardioprotective agent (e.g., dexrazoxane)[104,122].

A high cumulative dose of anthracycline is the greatest risk factor for AIC; however there is no absolutely safe dose for these chemotherapeutics, and even low doses < 200 mg/m^2 can cause cardiac damage[123]. In addition to the cumulative dose, other risk factors increase the risk of cardiotoxicity, including female gender, younger (in children) and advanced (in adults) age at diagnosis, radiation therapy (especially in the mediastinal region), concomitant therapy (e.g., trastuzumab, cyclophosphamide, bleomycin and vincristine), and the presence of cardiac disease and comorbidities, such as diabetes or obesity[104]. However, some patients appear to be more vulnerable than others, independent of these risk factors, and some others are not going to

experience cardiotoxicity; suggesting a genetic predisposition to AIC. Several studies[124–139] have identified genetic variants associated with AIC; including common variants in genes involved in anthracycline transport and metabolism (e.g., *CBR1* and *CBR3*[125,127], *SLC28A3*[133,134], *ABCB1*[133], *ABCC1*[129,133]) or pharmacodynamic genes (e.g., *NCF4*[124,125,131], *NQO1*[125]). Only Aminkeng et al.[139] used a genome-wide approach to identify susceptibility genetic variants to AIC. All the remaining studies are hypothesis-driven candidate-gene studies and all have focused on common genetic variants [minor allele frequency (MAF)≥5%]; thus, it is plausible that analyses of low-frequency (MAF=1-5%) and rare (MAF<1%) potentially functional variants could explain additional interpatient variability in susceptibility to AIC. Moreover, genetic studies may help to identify genomic biomarkers predictive of AIC and could help to elucidate the precise molecular mechanisms. In this thesis we focused on the identification of additional genes and susceptibility variants to chronic AIC, with particular attention to rare variation, by using exome array analysis in both, anthracycline-treated pediatric and adult oncology patients.

Table I3. Summary of the studies

Trait	Clinical problem	Aim	Study
Ewing sarcoma	• Rare (2.5-3 cases per million) and aggressive pediatric bone tumor • High incidence of recurrence (>40%) • Dismal outcome for patients with metastatic disease (cure rates=20-40%) and for those with recurrent/refractory disease (long-term survival <15%)	Identification of predictive and prognostic genetic variants	**Study I**
Capecitabine-induced hand-foot syndrome (CiHFS)	• Frequently used for the treatment of breast and colorectal carcinomas, especially for metastatic tumors • CiHFS is a common (>30%) cutaneous adverse event • ~17% of patients experience severe CiHFS (grade 3) resulting in dose reduction or therapy discontinuation and the inability to perform daily routines	Identification of additional genetic variants predictive of susceptibility to CiHFS apart from those involved in the enzymatic metabolic pathway of capecitabine biotransformation	**Study II**
Anthracycline-induced cardiotoxicity (AIC)	• Widely used chemotherapeutic agents in children and adults • Their clinical use is compromised by cardiotoxicity. • Children cancer survivors are particularly vulnerable to AIC • Symptomatic chronic AIC is rare (7.3% of patients), but subclinical alterations are quite frequent (up to 57% of patients) • Risk of cardiac events can persist several decades after therapy completion and can progress to fatal heart failure	Identification of novel genes and variants predictive of susceptibility to AIC with special attention to low frequency and rare coding variants	**Study III and IV**

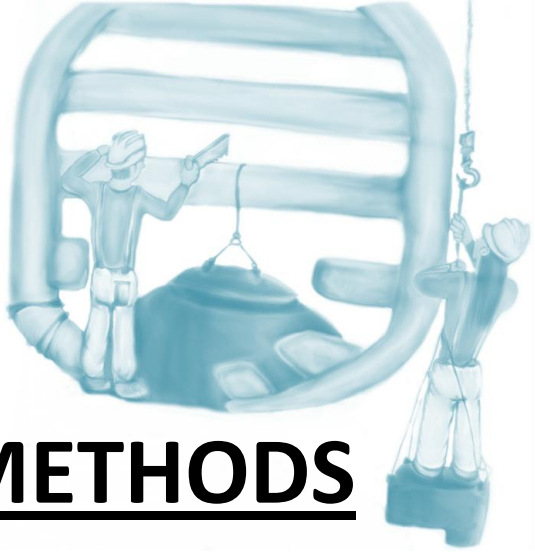
OBJECTIVES

The principal purpose of this thesis was to identify predictive and prognostic germline genetic variants for chemotherapeutics efficacy and toxicity, to provide essential data that will help to understand the molecular mechanisms underlying drug response and with the ultimate goal of improving anticancer treatment and establish the basis for an individualized therapy management in cancer patients. To achieve this, we developed the following specific aims:

1. To discover genomic biomarkers that predict treatment outcome in children diagnosed with Ewing sarcoma.

2. To identify novel genomic biomarkers associated with increased susceptibility to capecitabine-induced hand-foot syndrome (CiHFS) in adult breast and colorectal cancer patients.

3. To establish the contribution of low-frequency and rare genetic variants to susceptibility to chronic anthracycline-induced cardiotoxicity (AIC) in pediatric oncology patients and in adult breast cancer patients.



MATERIALS & METHODS

In order to facilitate the understanding of this section, detailed information about the materials and methods used in **Study I** and **Study II** is given separately. **Studies III** and **IV** used the same methodology, and so the materials and methods used are presented together. In addition, a table is provided summarizing the main materials and methods for each study.

1. Materials & Methods, Study I: identification of predictive and prognostic genetic variants for Ewing sarcoma

1.1. Patients

Eligible patients had histologically confirmed Ewing sarcoma diagnosed before age 30 years. The discovery cohort consisted of 106 Spanish Ewing sarcoma pediatric patients recruited between 1993 and 2012 at the *La Paz* University Hospital and the *Niño Jesús* University Hospital in Madrid and at the University Clinic of Navarra in Pamplona. The replication cohort consisted of 389 Ewing sarcoma pediatric patients from Austria (153), France (110), Italy (97) and Germany (29), recruited in 1991-2010, 1999-2007, 1987-2010 and 1994-2008, respectively. In both cohorts, patients were treated according to a multimodal protocol consisting of multiagent chemotherapy mostly involving combinations of vincristine (V), ifosfamide (I), doxorubicin (D), cyclophosphamide (C), etoposide (E) and/or actinomycin-D (A), combined with surgery and/or radiation therapy. In the induction chemotherapy, three main protocols were used: VIDE (62% of patients), VDC (17%) and VDC+VAI+VDC+IE (14%). Information about the neoadjuvant protocols is shown in **Table MM1**.

Table MM1. Neoadjuvant treatment given to Ewing sarcoma patients

Neoadjuvant chemotherapy	Discovery (N=106)	Replication (N=389)			
	Spain (106) N (%)	Germany (29) N (%)	Italy (97) N (%)	Austria (153) N (%)	France (110) N (%)
VIDE	27 (25%)	15 (52%)	-	153 (100%)	110 (100%)
VDC	79 (75%)	-	3 (3%)	-	-
VDC+VAI+VDC+IE	-	-	70 (72%)	-	-
VDI+C*E*+VDI+C*E	-	-	15 (15%)	-	-
VDCA+I	-	-	5 (5%)	-	-
VDIA	-	4 (14%)	-	-	-
EVDIA	-	4 (14%)	-	-	-
VDIA+EVDIA		1 (3%)	-	-	-
VIDE+VAI	-	3 (10%)	-	-	-
VIDE+VAI+VAC	-	1 (3%)	-	-	-
Other	-	1 (3%)	-	-	-
Missing	-	-	4 (4%)	-	-

Abbreviations: A, actinomycin-D; C, cyclophosphamide; C*, high-dose cyclophosphamide; D, doxorubicin; E, etoposide; E*, high-dose etoposide; I, ifosfamide; V, vincristine. Other: chemotherapy regimen involving vincristine, doxorubicin, etoposide and cisplatin.

Postoperative chemotherapy typically consisted of the administration of the VAC or the VAI regimen. Of patients in the replication cohort with metastasis at diagnosis and/or non-resectable primary tumor, 18% were treated with high-dose chemotherapy. None of the patients in the discovery cohort received high-dose chemotherapy.

Information on age at diagnosis, sex, primary tumor site, therapy, existence of metastasis at diagnosis and development of recurrence was abstracted from medical records. Where possible, tumor response to treatment, defined as the percentage of necrosis induced in the tumor after neoadjuvant chemotherapy, was determined histologically. Recurrence was defined as any evidence of new disease during/after the completion of therapy, including both locoregional and distant disease relapses. Overall survival was calculated as the time from tumor diagnosis until death from any cause or date last known to be alive.

Written informed consent was obtained from adult patients and from the parents or legal guardians of children. The study was approved by the ethics committees of all participating universities and hospitals.

1.2. Selection of genes and polymorphisms

Since often transporters and metabolizing enzymes are shared between different drugs and the existence of a functional interplay between them in drug absorption and disposition, it seems essential the integration of multiple drug pathways to allow a comprehensive analysis of genetic factors influencing pharmacokinetics and drug efficacy, especially when polychemotheraphy is administered. Therefore, we decided to adopt a genotyping-based candidate pathway strategy and selected 24 genes reported to be involved in the pharmacokinetics of the 6 agents (vincristine, ifosfamide, doxorubicin, cyclophosphamide, etoposide and actinomycin-D) commonly used in chemotherapy regimens for Ewing sarcoma, based on the literature and on the information available in the Pharmacogenomics Knowledge PharmaGKB database[140] (**Table MM2**).

Table MM2. Candidates genes studied

Category	Genes
Transporters	*ABCA3, ABCB1, ABCC1, ABCC2, ABCC3, ABCC4, ABCC6, ABCG2, SLC31A1, SLCO6A1, SLC19A1*
Phase I metabolism enzymes	*MPO, SOD1, ALDH1A1, CYP3A4, CYP3A5, CYP2A6, CYP2B6, CYP2C8, CYP2C9, CYP2C19*
Phase II metabolism enzymes	*GSTM1, GSTP1, GSTT1*

A total of 384 common variants were selected across these candidate genes, as previously described[141]. Briefly, both SNVs with potentially functional effects (causing amino acid changes, potentially causing alternative splicing, located in the promoter region, in putative transcription factor binding sites, or disrupting miRNAs and their targets) identified using the bioinformatics tool PupaSuite[142], and other functional variants already described in the literature were selected. In addition, across these genes we selected tagSNVs to ensure a more extensive evaluation of the contribution of common variation within each gene using Haploview software (v.4.0)[143]. MAF>5% and r^2 threshold of 0.8 were employed for tagging procedure. The preliminary list of variants was filtered using as criteria suitability for the Illumina genotyping platform (selecting only those with an assay score >0.6, associated with a high success rate) and MAFs of at least 5%. A final number of 384 variants relevant to this study were included in an oligonucleotide pool assay for analysis using the Illumina Veracode technology (Illumina, San Diego, USA).

1.3. Isolation and quantification of DNA

Five to ten ml of peripheral blood samples were collected in tubes containing anticoagulant. Whole blood was fractionated by centrifuging during 10 min at 3000 rpm; the plasma upper layer was aspirated off, while the buffy coat containing the white blood cells was extracted by a Pasteur pipette. Germline DNA was extracted using standard phenol-chloroform extraction protocols and an automatic DNA extraction instrument (Magnapure, Roche, Mannheim, Germany). DNA was quantified using the Quant-iT™ PicoGreen dsDNA kit (Invitrogen, Carlsbad, USA). For the standard curve, a series of dilutions of genomic DNA (Clontech, Mountain View, USA), giving a final DNA concentration from 20 to 200 ng/µl, were prepared in TE buffer (10 mM Tris, 1 mM EDTA, pH 7.5). The standards and 2 µl of each sample were pipetted into a 96 well microplate (Falcon, BD Biosciences, San Jose, USA). PicoGreen reagent was diluted in TE buffer according to the kit instructions and 198 µl of the mix was pipetted in the wells. The fluorescence was read at 520 nm after 480 nm excitation using the DTX 800 Multimode Detector (Beckman Coulter, Fullerton, USA).

250 ng of genomic DNA from each sample were genotyped using the GoldenGate assay with VeraCode technology (Illumina, San Diego, USA) on the BeadXpress platform according to the published protocol. Illumina GoldenGate technology is based on allele-specific primer extension assays (**Figure MM1**). Briefly, in the first step DNA is activated through a chemical reaction with biotin to enable binding to streptavidin-conjugated paramagnetic particles and then hybridized with the assay oligonucleotides. Three oligonucleotides are designed to query the allele at each

SNV locus. Two oligos are specific to each allele of the SNV site, called the Allele-Specific Oligos (ASOs). Each ASO contains a region of genomic complementarity but has a different nucleotide on the 3' end, which corresponds to the complement of each possible allele in the genomic DNA; and also contains a unique universal PCR primer sequence. A third

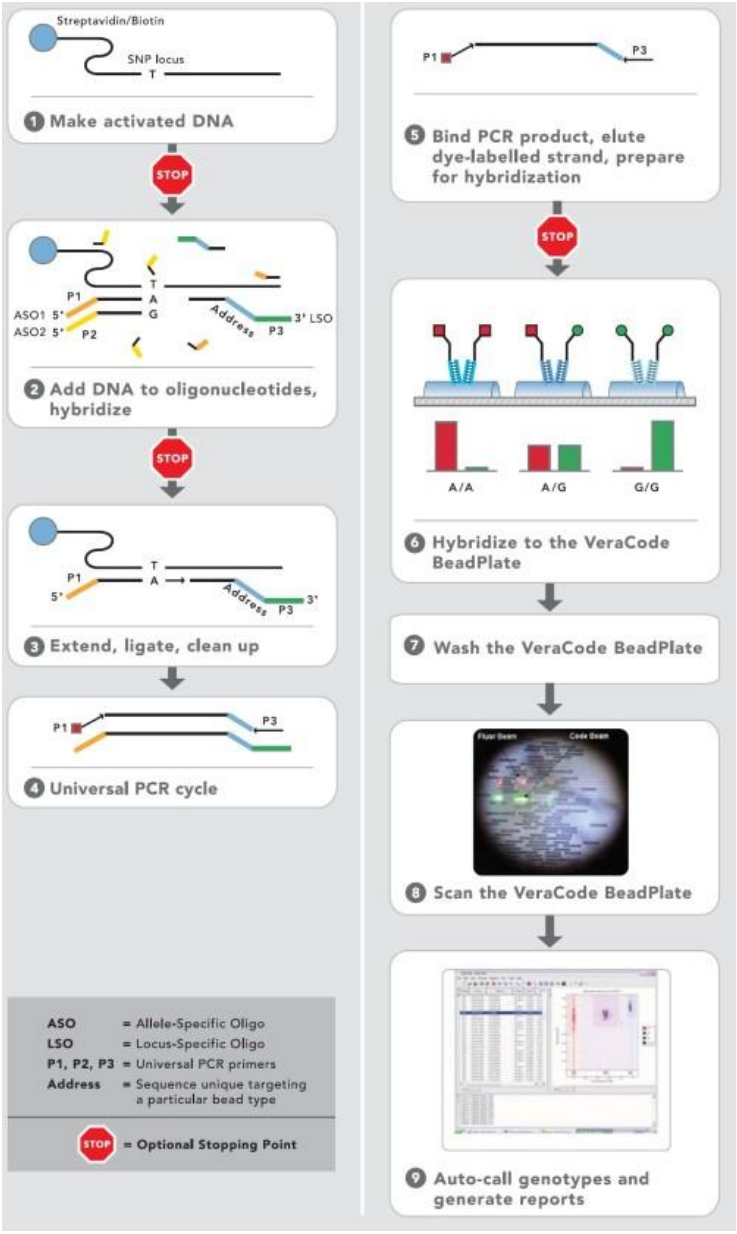


Figure MM1. Illumina VeraCode GoldenGate protocol (from Illumina website[144])

oligo, the Locus Specific Oligo (LSO), hybridizes several bases downstream from the SNV site and contains an additional unique address sequence that targets a particular VeraCode bead type and also contains a region of genomic complementarity and a universal PCR primer site. After the oligo hybridization, allele-specific extension of the appropriate ASO and ligation of the extended product to the LSO were performed. These joined full-length products provide a

template for PCR using universal PCR primers P1, P2, and P3. Universal primers P1 and P2 are Cy3 and Cy5 labeled to allow allele discrimination. PCR products were hybridized to their VeraCode bead type through their unique address sequence and each VeraCode microtitre bead plate was imaged on the Illumina BeadXpress Reader System (Illumina, San Diego, USA). After scanning, SNV genotype clustering and individual sample genotype calls were interrogated using the Illumina GenomeStudio software (v.3.2)[144].

We excluded variants with a call rate <0.95 with MAF<0.05, whose genotype distribution deviated from Hardy-Weinberg equilibrium ($P<10^{-6}$), with Mendelian allele-transmission errors, or with discordant genotypes between duplicate samples. Samples with a call rate <0.90 were excluded.

1.4. Statistical analysis

We studied the association of SNVs with tumor response and overall survival. Patients were divided into two categories: good responders, with tumor necrosis ≥90%; and poor responders, with tumor necrosis <90%[145]. Odds ratios (ORs) and 95% confidence intervals (CIs) for good tumor response by genotype were estimated using logistic regression analysis. Variants for which associations with $P<0.05$ were observed were assessed in the European replication cohort.

Table MM3. Clinical information recorded from Ewing sarcoma patients and associations in univariable analyses with tumor response and overall survival

	Tumor response		Overall survival	
	Discovery	Replication	Discovery	Replication
Clinical information	**P**	**P**	**P**	**P**
Age at diagnosis	**0.039**	**0.002**	**0.048**	**0.038**
Sex	0.52	0.38	0.55	0.79
Primary tumor site	0.29	0.99	0.33	0.69
Metastasis at diagnosis	**0.001**	**0.042**	**1.67×10^{-6}**	**6.8×10^{-12}**
Neoadjuvant therapy	0.55	0.54	0.15	0.15
Response to treatment	NA	NA	**0.029**	**6.0×10^{-5}**
Recurrence	0.093	0.082	**1.67×10^{-6}**	**8.7×10^{-22}**
Vital status	NA	NA	NA	NA
Overall survival	NA	NA	NA	NA
Country of origin	NA	0.90	NA	0.81

Country of origin was included as covariate in analyses with overall survival in the replication cohort due to incomplete information regarding adjuvant therapy protocols. Abbreviations: NA, not applicable.

We also tested associations between SNV genotypes and overall survival using Cox-regression analysis. Variants with $P<0.05$ in the discovery set were assessed in the replication cohort.

Survival curves were plotted using the Kaplan-Meier product-limit method, and the significance of differences between these curves was determined using the log-rank test.

Clinical factors with associated $P<0.05$ in univariable analyses (**Table MM3**) with tumor response or overall survival were included as covariates in corresponding multivariable analyses.

In addition to the additive genetic model, we considered dominant and recessive models.

Analysis were carried out using PLINK (v.1.07)[146,147] or SPSS software (v.18.0; SPSS, Chicago, USA).

1.5. Functional annotations

We used information from the Encyclopedia of DNA Elements (ENCODE)[148] using custom tracks on the UCSC Genome browser[149] and HaploReg[150] to investigate whether the risk-associated SNVs or their correlated SNVs [$r^2 \geq 0.8$] had potential regulatory functions. ENCODE describes genes, transcripts, and transcriptional regulatory regions, as well as DNA binding proteins, that interact with regulatory regions in the genome, including transcription factors, histones and other markers, and DNA methylation patterns that define states of the genome in various cell types[148]. HaploReg is a web tool that provides information about predicted chromatin states in nine cell types, conservation across mammals and effects on regulatory motifs[151].

Table MM4. Summary of the main materials and methods of Study I

Patients	**Discovery cohort:** 106 Spanish pediatric Ewing sarcoma patients **Replication cohort:** 389 pediatric Ewing sarcoma patients from across Europe
Genotyping	**Candidate pathway approach:** analysis of 24 pharmacokinetic genes involved in the transport or metabolism of the 6 agents commonly used for Ewing sarcoma treatment (Illumina VeraCode GoldenGate). 384 SNVs were selected across these 24 candidate genes.
Statistical analysis	**Single-variant associations:** • Tumor response: logistic regression analysis • Overall survival: Cox-regression analysis • Significance level: $P<0.05$
Functional annotations	ENCODE (UCSC Genome browser) and Haploreg

2. Materials & Methods, Study II: identification of genetic variants predictive of susceptibility of capecitabine-induced hand-foot syndrome (CiHFS)

2.1. Patients

Capecitabine-treated adult breast and colorectal cancer patients were recruited retrospectively through several Hospital Oncology Units across Spain. The discovery cohort consisted of 166 patients recruited at the *San Carlos* University Hospital (Madrid, Spain) (N=87) and the *Virgen de la Victoria* University Hospital (Málaga, Spain) (N=79). The validation cohort consisted of 85 patients recruited at the *Gregorio Marañón* University Hospital (Madrid, Spain) (N=58) and 27 patients enrolled in a clinical trial comparing standard versus continuous administration of capecitabine in metastatic breast cancer (www.clinicaltrials.gov, identifier code: NCT00418028). To ensure accurate patient phenotyping and maximize genotyping efficiency we applied an extreme phenotype sampling[152]: case patients selected for both discovery and replication studies were required to have grade 3 toxicity experienced across any treatment cycle before completing eighteen cycles of treatment (over one year) and control patients were those not developing CiHFS during treatment (grade 0). Patients who did not fulfill these criteria were excluded. CiHFS grade was determined according to the NCI Common Terminology Criteria for Adverse Events v 4.0 (CTCAE v4.0)[86]. Cumulative dose was calculated as the amount in mg/m^2 of capecitabine taken prior to the development of toxicity or the end of follow-up, whichever was lower. Capecitabine was administered according to different schedules. Colorectal cancer patients were treated with a standard regimen (1250 mg/m^2 orally every 12 h on days 1-14, every 3 weeks), while breast cancer patients were treated either with the same standard regimen or with a continuous regimen (800 mg/m^2 orally every 12 h daily during 21 days). All protocols were approved by the institutional review boards at each institution, and all subjects provided written informed consent.

Information on age at diagnosis, sex, primary tumor, treatment regimen, capecitabine cumulative dose and CiHFS grade was recorded from medical records

2.2. Isolation and quantification of DNA

Genomic DNA was isolated from peripheral blood lymphocytes and quantified using the same procedures above described (section 1.3).

2.3. Genotyping

250 ng of genomic DNA from each sample in the discovery cohort were genotyped using Illumina Human610-Quad BeadChips containing 616,795 variants according to the manufacturer's protocols (Illumina, San Diego, USA). The Illumina Human610-Quad array is based on the Infinium assay (Illumina, San Diego, USA) (**Figure MM2**).

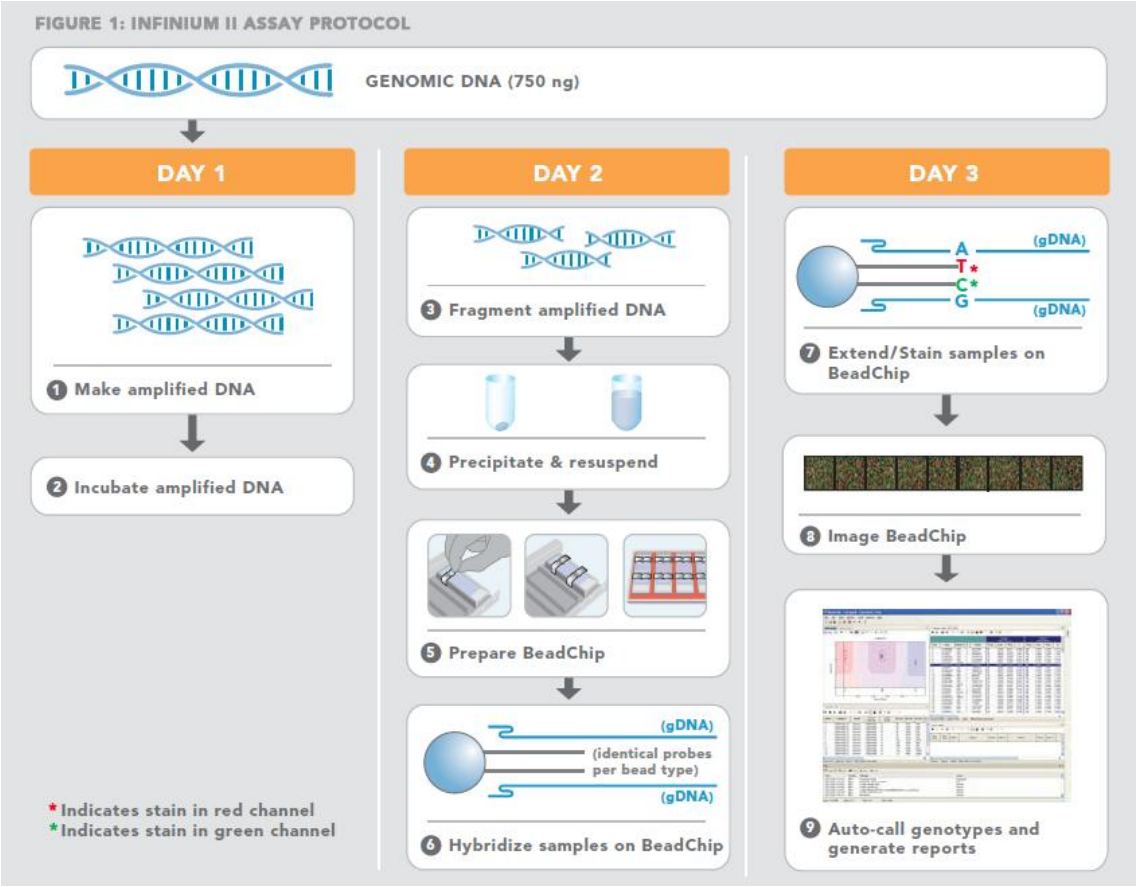


Figure MM2. The Illumina Infinium protocol (from Illumina website, www.illumina.com)

In this system a whole-genome amplification step is used to increase the amount of DNA up to 1000-fold. The DNA is fragmented and captured on BeadChips by hybridization to immobilised SNV-specific primers, followed by single-base primer extension. Single-base extension of the oligos on the BeadChip, using the captured DNA as a template, incorporates detectable labels on the BeadChip and determines the genotype call for the sample. If there is a perfect match, extension occurs and signal is generated. If there is a mismatch, extension does not occur and no signal is generated. BeadChips are imaged using the Illumina iScan System, a two-channel high-resolution laser imagers that scan BeadChips at two wavelengths simultaneously and create an image file for each channel (e.g., two per array). The GenomeScan software determines intensity values for each bead type and creates data files for each channel. The GenomeStudio software

package extracts whole-genome DNA analysis data from image data files created by the Illumina BeadArray Reader[153].

DNA samples from the replication cohort were genotyped by KASpar Assays (Kbioscience, UK) and fluorescence was determined by the sequencer Detection System 7900HT (Applied Biosystems, Foster City, USA) (**Figure MM3**).

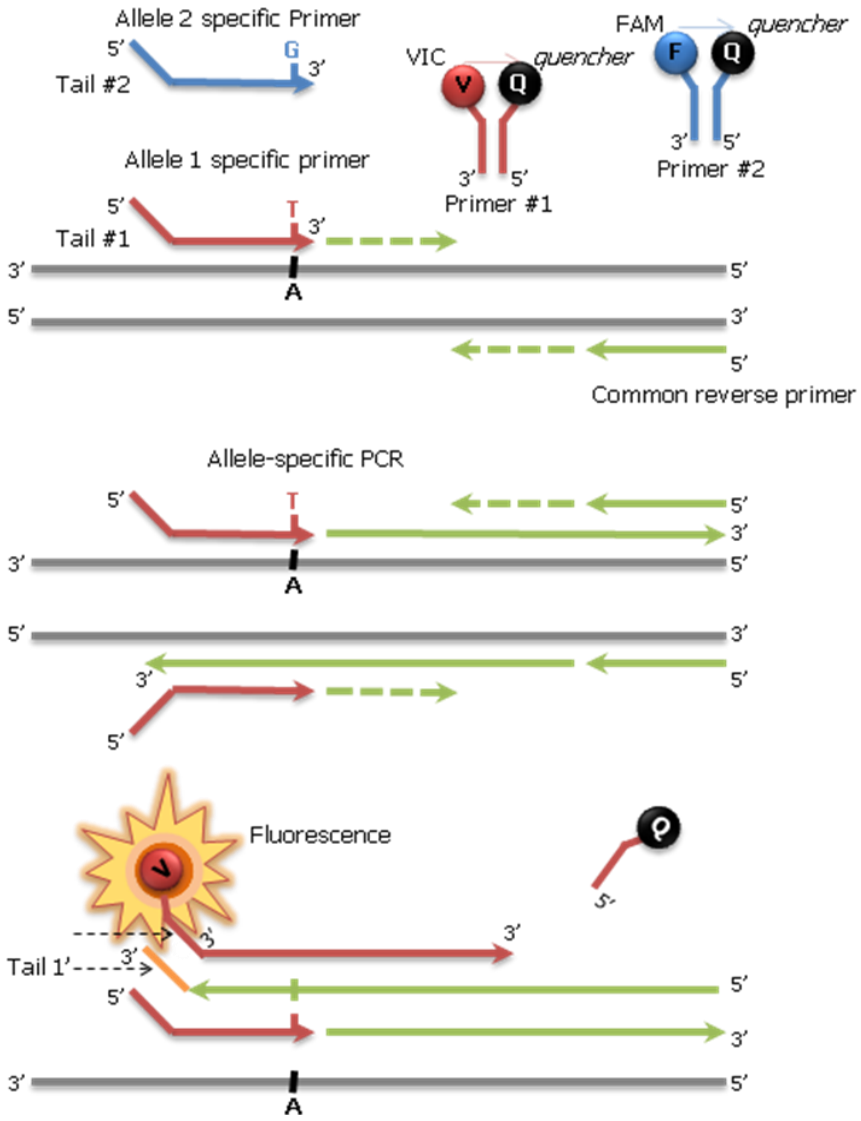


Figure MM3. KASPar SNV genotyping system (adapted from http://www.lgcgroup.com/services/KASP_genotyping).

The KASpar genotyping system is based on:

- two allele-specific primers (one for each SNV allele). Each primer has an unlabelled tail sequence at the 5' end and the two primers have different tail sequences.
- a common reverse primer.

- two fluor-labelled oligos complementary to the sequences of the tails of the allele-specific primers. This oligos are labelled with a fluorochrome (FAM or VIC) and with a quencher; so that there is no fluorescence emission.

In the beginning of the PCR, the appropriate allele-specific primer and the common primer binds to the DNA at the SNV locus and PCR occurs. As PCR proceeds further, the fluor-labelled oligo becomes incorporated into the template as well, and is hence no longer bound to its quencher-labelled complementary oligo. As the fluorescence is no longer quenched, the appropriate fluorescent signal is generated and detected by the usual means (**Figure MM3**). If the genotype of a diploid individual for a particular SNV is homozygous, only a FAM or VIC signal will be generated. If the individual is heterozygous, a FAM and VIC signal will be generated. The Sequence Detection System (SDS) Software (v.2.2.2) uses the fluorescence measurements made during the plate read to plot fluorescence values based on the signals from each well. The plotted fluorescence signals indicate which alleles are in each sample (**Figure MM4**). Coriell sample/s in duplicate (Coriell Cell Repository, Camden, USA) as positive controls (e.g., DNAs with known genotypes) and NTC (no template control) were included in all assays.

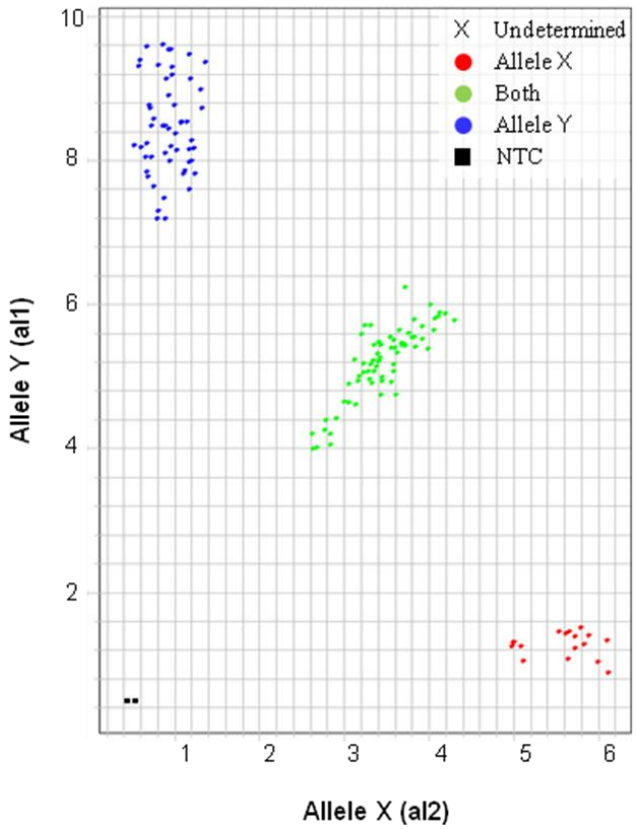


Figure MM4. KASPar allelic discrimination. Each data point represents the fluorescence signal of an individual DNA sample. Samples of the same genotype will have generated similar levels of fluorescence and will therefore cluster together on the plot. Abbreviations: NTC, no template controls.

Fine-mapping genotyping was performed by using the MassARRAY genotyping system (Sequenom, San Diego, USA) and Infinium assay (Illumina, San Diego, USA) following the manufacturer's instructions. The MassARRAY genotyping system consists of an initial locus-specific PCR reaction, followed by a locus-specific primer extension reaction (iPLEX assay) in which an oligonucleotide primer anneals immediately upstream of the polymorphic site being genotyped. In the iPLEX assay, the primer and amplified target DNA are incubated with mass-modified dideoxynucleotide terminators. The primer extension is made according to the sequence of the variant site, and is a single complementary mass-modified base. Through the use of MALDI-TOF mass spectrometry, the mass of the extended primer is determined. The primer's mass indicates the sequence and, therefore, the alleles present at the polymorphic site of interest[154] (**Figure MM5**).

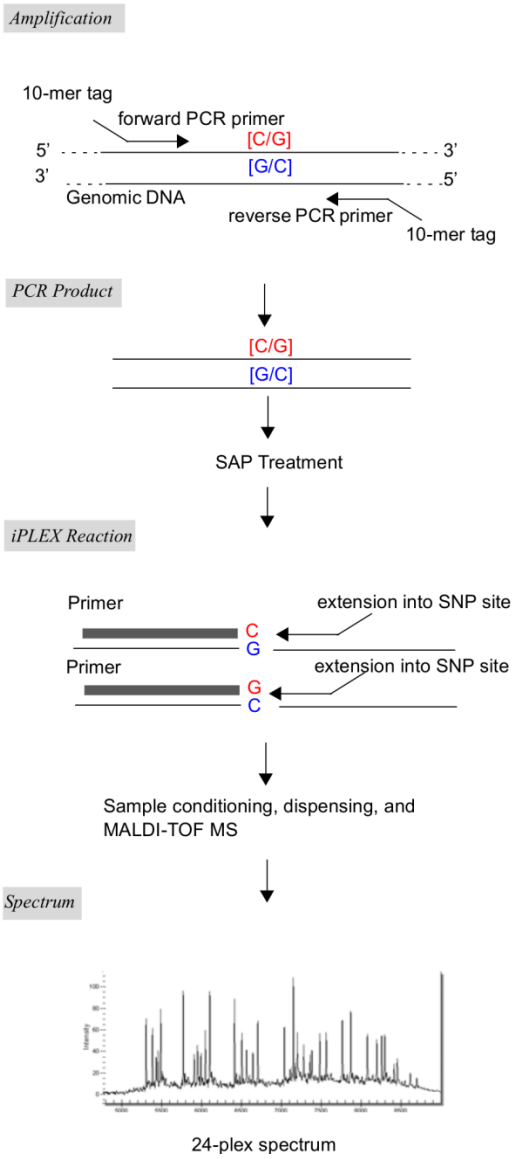


Figure MM5. Genotyping using the Sequenom MassARRAY iPLEX platform[154]. Abbreviations: SAP, Shrimp Alkaline Phosphatase; MS, mass spectrometry.

2.4. Quality control (QC)

Samples from the discovery cohort were excluded from further analyses if they had more than 5% missing genotype. Variants excluded were non-diploid variants, CNVs and SNVs deemed unreliable by Illumina (Tech Note: Infinium Genotyping Data Analysis, 2007), variants with missing genotypes in more than 5% of samples, with MAF <0.05 as well as SNVs whose genotype frequencies departed from Hardy-Weinberg equilibrium at $P<10^{-6}$. To control for potential population stratification, we performed a principal component analysis (PCA)[155] using the snpStats package in R (v.3.0.1) considering only variants that passed QC (520,052).

2.5. Statistical analysis

Genotypes were assessed in relation to the cumulative dose of capecitabine using Cox regression analysis; modeling the cumulative dose of capecitabine (mg/m^2) up to the development of at least grade 3 CiHFS for cases. Patients not experiencing CiHFS of any grade during capecitabine treatment (controls) were censored at the total cumulative dose received. Estimated hazard ratios (HRs) were interpreted as relative risks of developing severe CiHFS at a given cumulative dose. Cox regression analysis was performed in the global data set [discovery cohort (166) + validation cohort (85), N=251] to identify clinical factors associated with CiHFS development. Associations were assessed under an additive genetic model by multivariable analysis, adjusting for sex, tumor type and treatment regimen. Cox regression analyses were performed using the R statistical environment (v.3.0.1) and PLINK (v.1.07)[146,147]. Kaplan-Meier comparisons of cumulative dose of capecitabine taken up to the development of grade 3 CiHFS according to SNV genotype were performed using the SPSS software (v.18.0; SPSS, Chicago, USA) and the log-rank test was applied.

A quantile-quantile (Q-Q) plot comparing the distribution of *P*-values observed (from Cox regression analysis adjusted for sex, tumor type and treatment regimen for 520,052 variants that passed QC in the discovery cohort) to those expected to account for population stratification was generated using the qqman package in R (v.3.0.1). The genomic inflation factor (λ) based on Q-Q plot was calculated using R (v.3.0.1).

Variants with associated *P*-values below a predefined threshold of 10^{-5} in the discovery cohort were assessed using the same statistical methods.

Haplotype estimation for the four most strongly associated SNVs ($P<5x10^{-7}$ and MAF>5%) was performed using PHASE[156] (v2.1.1) and haplotype-specific HR, *P*-values and confidence limits were estimated using SPSS software (v.18.0; SPSS,Chicago, USA).

2.6. Block definition, variants selection and imputation

The boundaries for the linkage disequilibrium (LD) block containing the replicated SNV rs6093063 in 20q13.33 locus were defined by the furthest upstream and downstream correlated SNVs with rs6093063 ($r^2>0.10$). Based on data from the June 2012 release of the 1000 Genomes Project[21], we selected for genotyping all SNVs with $r^2>0.10$ with the marker plus a minimum set of SNVs tagging all other SNVs in the block at $r^2>0.8$, the latter identified using Haploview software (v.4.2)[143]. We catalogued 605 variants at the region and we successfully genotyped 34. Prediction of the untyped SNVs within the LD block was carried out using IMPUTE (v.2.0)[157] based on phased haplotypes from 1000 Genomes Project data[21]. Genotypes for 80 SNVs were reliably imputed (imputation information score>0.3).

2.7. Cell lines and tissue samples

Cell lines: HaCaT cells (spontaneously immortalized human keratinocytes) were provided by Manuel Serrano (Spanish National Cancer Research Centre, CNIO, Madrid, Spain) and cultured in MEME-based (M8167, Sigma) low calcium medium (Nowak and Fuchs, 2009) containing 0.05 mM Ca^{2+}. Normal Human Epidermal Keratinocytes (NHEK) were purchased from Lonza (Cat Nº: 00192627; Lot Nº: 0000246915) and grown in Lonza's KGM-Gold media (basal medium supplemented with bovine pituitary extract, human epidermal growth factor, bovine insulin, hydrocortisone, Gentamicin, Amphotericin-B, Epinephrine and Transferrin; Cat Nº 00192060). Proliferative keratinocytes are normally found in the basal layer of the epidermis and move to suprabasal layers upon differentiation. In culture, keratinocytes proliferate and grow in a monolayer in low calcium medium, but the addition of calcium promotes cell-cell contact formation and differentiation into post-mitotic keratinocytes[158]. So, to induce keratinocyte differentiation, 2 mM Ca^{2+} was added to the medium. Cell lines were maintained under standard conditions and routinely tested for Mycoplasma.

Tissue samples: fifty healthy liver tissue samples were obtained from the Spanish National Cancer Research Center (CNIO, Madrid, Spain) Tumor Bank. The study was approved by the ethics committee of the Spanish National Cancer Research Center (CNIO, Madrid, Spain).

2.8. Circular chromosome conformation capture (4C)-sequencing

Human epidermal keratinocytes, NHEK and HaCaT, homozygous for the reference or the risk haplotypes, respectively, were treated with 2 mM Ca^{2+} for 72 h before collection. Preparation of 4C samples was performed as previously described[159] with some modifications (**Figure MM6**). Briefly, $2x10^7$ cells of each haplotype were trypsinized, PBS (phosphate buffered saline) washed and resuspended in 10 ml of fixing solution containing 1% formaldehyde in PBS. After 10 min, 0.125 M glycine was added to stop fixation and the mixture was incubated for 5 min at room temperature. Cells were pelleted and nuclei isolated after 40 min of incubation in cold lysis buffer (50 mM Tris pH 7.5, 50 mM NaCl, 5 mM EDTA, 1% NP40, 2% TX-100 + protease inhibitors cocktail). Proper lysis was determined by methyl green-pyronin staining. HindIII and DpnII (NEB) were used as first and second cutters, respectively. 4C libraries were amplified using long primers with 18-20 bp homology to the bait sequence and Illumina paired-end adapter flanks. Primer sequences were chosen to viewpoint sites, which were within the *CDH4* promoter region as follows:

P5_Vp1:AATGATACGGCGACCACCGAACACTCTTTCCCTACACGACGCTCTTCCGATCTCTACAGATGTTTGCTAAGCTT

P7_Vp1: CAAGCAGAAGACGGCATACGAAGAATCAATGTAGCAGGTCT

P5_Vp2:AATGATACGGCGACCACCGAACACTCTTTCCCTACACGACGCTCTTCCGATCTGCCCCCTCCCAGAAGCTT

P7_Vp2: AAGCAGAAGACGGCATACGAATTTCCGAGGTTACAACAGCGC

4C-seq data analysis and normalization was performed with 4Cseqpipe[159] on the Genome Reference Consortium Human Build 37 (GRCh37/hg19).

2.9. Functional annotations

We used information from ENCODE[148] using custom tracks on the UCSC Genome browser[149] to investigate whether the risk-variant containing locus might have potential regulatory functions.

2.10. mRNA expression analysis in nontumoral human liver tissues

Total RNA was extracted from cryopreserved liver tissues using 1ml TriZol (Invitrogen, CA, USA) and treated with DNase I (Ambion, Austin, USA) according to the instructions of the manufacturer. 1 µl of RNA was used to measure the concentration by Nanodrop ND-1000 (Wilmington, DE, USA) and the RNA quality was tested through 1% agarose gel electrophoresis.

Reverse transcription was performed in a single tube using 1 µg of the total RNA, Superscript II Reverse Transcriptase (Invitrogen, CA, USA) and oligo random hexamers primers following the manufacturer's instructions.

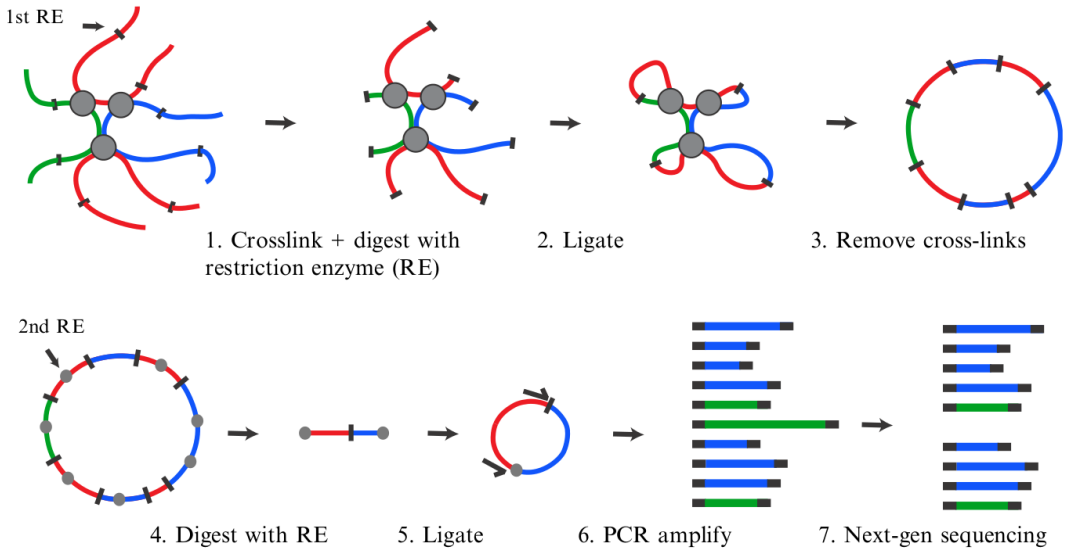


Figure MM6. Outline of the 4C technology[159]**.** After cross-linking by formaldehyde and digestion with the first restriction enzyme (RE), "hairballs" of cross-linked DNA are created (**1**). Chromatin is diluted and religated to fuse the ends of DNA fragments present in the same "hairball" (**2**). The ultimate outcome of this ligation event is large DNA circles encompassing multiple restriction fragments. Cross-links are removed by heating (**3**); DNA is digested by a second RE (usually a four cutter) (**4**) and religated under diluted conditions to create small DNA circles, most of which carry a primary ligation junction. Inverse PCR primers specific for the fragment of interest ("viewpoint") and carrying 5´adapter overhangs for next-generation sequencing allow amplification of all its captured sequences followed by high-throughput sequencing. Abbreviations: Next-gen, next generation.

CDH4 mRNA was quantified by real-time quantitative PCR (qRT-PCR) with the ABI PRISM® 7900HT Sequence Detection System (Applied Biosystems, Foster City, USA), using human *CDH4*-specific primers (**Table MM5**) and fluorescent probes from the Human Universal ProbeLibrary Set (Roche, Mannheim, Germany). The expression levels of PCR products were normalized according to *ACTB* gene expression (**Table MM5**). The final concentrations of the primers and the probes were 500 nM and 1000 nM, respectively, with a final volume of 12 ul, using 2.4 ul of a 1:10 cDNA dilution for each reaction. Negative controls were present in all series of qRT-PCR reactions and all assays were carried out in triplicate. For the amplification reaction, the Universal PCR Master Mix (PE Applied Biosystems, Foster City, USA) was used and the conditions consisted in an initial step of 95 ºC for 10 min, followed by 55 cycles of 15 s at 95 ºC and 1 min at 60 ºC. The delta–delta Ct method[160] was used for the calculation of the different amounts of mRNA.

Association between the risk locus haplotypes and expression levels of *CDH4* gene was assessed using the Student's t test, after natural log transformation to obtain normally distributed expression data. Differences were considered significant when $P<0.05$. These analyses were carried out using SPSS (v.18.0; SPSS, Chicago, USA).

Table MM5. *CDH4* and *ACTB* specific primers

Gene	Forward Primer	Reverse Primer	Probe (5´-3´)
CDH4	AGCACGCCTCTTACCACCT	ACGTAGATGTACAGGTCGATGG	UPL #39[a]
ACTB	CCAACCGCGAGAAGATGA	CCAGAGGCGTACAGGGATAG	UPL #64[a]

a. Human Universal ProbeLibrary Reference Gene Assay

2.11. R-cadherin and involucrin protein expression in skin samples

Normal paraffin-embedded skin samples derived from mastectomy of 16 breast cancer patients before capecitabine treatment (12 CiHFS grade 3 samples and 4 CiHFS grade 0 samples) were available for tissue microarray (TMA) construction. Duplicate 2-mm tissue cores using the MTA1-Manual Tissue Arrayer (Beecher Instruments) were selected from each skin sample. 3 µm thick tissue sections were cut from TMAs for R-cadherin and involucrin expression analysis by immunohistochemistry (IHC). Further IHC analyses of additional cadherins (E-cadherin, P-cadherin and N-cadherin) were performed to confirm the R-cadherin antibody specificity and to compare the location of the different cadherins in the skin layers. IHC was done using the following protocol: after overnight incubation at 55°C and deparaffinization, antigen retrieval was performed using EnVision FLEX TRS High pH (Dako, Glostrup) for 20 min. Blocking of endogenous peroxidase activity, incubation with primary antibody, signal enhancement of R-cadherin and involucrin staining with EnVision FLEX HRP (Dako, Glostrup) and E, P and N-cadherins staining with EnvisionFlex+Mouse (Linker) (Dako, Glostrup) and revealing immunoreactivity were performed in an Autostainer Link 48 (Dako, Glostrup). The following primary antibodies were applied: R-cadherin: dilution 1:300, (HPA015613, Sigma, St. Louis); E-cadherin: ready to use (1R059, Dako, Glostrup); P-cadherin: dilution 1:75, (610228, BD Biosciences, New Jersey, USA); N-cadherin: dilution 1:100, (M3613, Dako, Glostrup); involucrin: ready to use (Thermo Fisher Scientific, CA, USA). Incubation times for the primary antibodies were 20 min (E-cadherin and P-cadherin) and 40 min (R-cadherin, N-cadherin and involucrin). Slides were counterstained with hematoxylin using an automatic slide stainer Varistain Gemini ES (Thermo Fisher Scientific, CA). Immunostained slides were digitized using a Zeiss Mirax automated slide scanner (Zeiss, Germany) with an objective of 40x magnification (0.12µm/pixel). Digital image analysis was performed by importing the Mirax file into the image analysis

software HistoQuant (3DHistech Ltd, Budapest, Hungary). For R-cadherin and involucrin expression quantification, the epidermis was selected according to morphological differences and IHC staining (**Figure MM7**). HistoQuant algorithm was used to quantify the intensity of R-cadherin and involucrin positive staining. Each positive area was analyzed at pixel level with each pixel containing staining intensity information corresponding to the red-green-blue (RGB) channels. In digital image analysis, the pixel intensity values for any color range from 0 to 255, wherein, 0 represents the darkest shade of the color and 255 represent the lightest shade of the color as standard. Since the optical density is proportional to the concentration of the stain, R-cadherin and involucrin positive staining was measured as the average of RGB channels and subtracted from 255 to associate large intensity values with a high R-cadherin or involucrin staining[161,162]. Differences in R-cadherin and involucrin expression were assessed using Student's t test and the SPSS software (v.18.0; SPSS, Chicago, USA). A *P*-value of <0.05 was considered to be statistically significant.

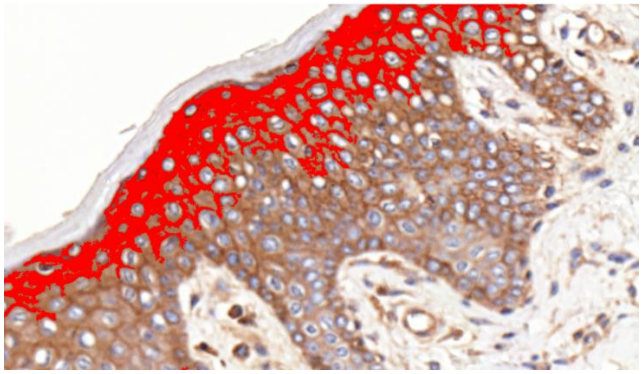


Figure MM7. Quantification of R-cadherin an involucrin expression in the skin from breast cancer patients sampled prior capecitabine treatment. For automated image acquisition, the TMA areas of interest were selected according to the epidermal morphological differences and brown marker (DAB) intensity. These image areas were used as training samples for the HistoQuant pattern recognition software to set quantification thresholds for the analysis. For that purpose the HistoQuant Wizard was used to set the algorithm, establishing different values for contiguous nucleus separation, cytoplasmic area definition, colour intensity separation for quantification and grading and size exclusion. For quantification of R-cadherin expression only the suprabasal layers of the epidermis were selected (in red). For quantification of involucrin, positive staining areas (corresponding to the cornified envelope of the epidermis) were selected. Each positive area was analyzed at pixel level with each pixel containing staining intensity information corresponding to the RGB channels.

2.12. R-cadherin expression knock-down (KD)

For knock-down experiments HaCaT cells were transfected with the short hairpin RNA (shRNA)expressing pGFP-V-RS plasmid (TR30007, OriGene) 12 h after plating using X-tremeGENE 9 DNA transfection reagent (Roche, Mannheim, Germany) according to the manufacturer's instructions. 24 h after transfection 2 mM Ca^{2+} was added to the medium for 48 h. Control cells

were transfected with a scramble shRNA expressing plasmid. The following R-cadherin shRNA sequences were used:

Clon 1: AGGCGACATCGGTGACTTCATCAATGAGG

Clon 2: TATGTTCACCATCAACAGCGAGACTGGAG

Antibodies: the following antibodies were used: β-actin mouse mAb (A5441, Clone AC-15, Sigma, St. Louis, USA), E-cadherin human mAb (Clone HECD-1, ab1416, Abcam), R-cadherin rabbit pAb (HPA015613, Sigma, St. Louis, USA), filaggrin rabbit pAb (PRB-417P, Covance, New Jersey, USA), loricrin rabbit pAb (PRB-145P, Covance, New Jersey, USA), involucrin rabbit pAb (PRB-140C, Covance, New Jersey, USA).

RNA isolation and qRT-PCR: total RNA was isolated from cells using TriZol reagent (Invitrogen, CA, USA) and treated with DNase I (Ambion, Austin, USA) according to the instructions of the manufacturer. 1 µl of RNA was used to measure the concentration by Nanodrop ND-1000 (Wilmington, DE, USA) and the RNA quality was tested through 1% agarose gel electrophoresis. 2µg were used for cDNA synthesis using the Ready to Go You Prime It First-Strand beads and random primers (GE Healthcare, Uppsala, Sweden). qRT-PCR reactions for *CDH1* (E-cadherin), *FLG* (filaggrin), *KRT1* (keratin 1), *KRT10* (keratin 10) and *IVN* (involucrin) were performed using the GoTaq qPCR Master Mix (Promega, Wisconsin, USA) in a MasterCycler Ep-Realplex thermal cycler (Eppendorf, Hamburg, Germany), with the following settings: 2 min at 95° C for initial denaturing and 35 cycles of 15 s at 95°C denaturing, 40 s at 57°C annealing, and 45 s at 72°C extension. The expression levels of PCR products were normalized according to *GAPDH* gene. Primers sequences for *CDH1, FLG, KRT1, KRT10, IVN* and *GAPDH* are shown in **Table MM6**.

qRT-PCR reactions for *CDH4* were carried out in an ABI PRISM 7900HT Sequence Detection System (Applied Biosystems, Foster City, USA), using the Hs00899698_m1 TaqMan Gene Expression probe and the Hs9999905_m1 TaqMan Gene Expression probe for the *GAPDH* gene, used as reference. 2.4 ul of a 1/10 CDH4 cDNA dilution were used for each reaction, and combined with 6 ul of 2X TaqMan Universal PCR Master Mix (No AmpErase UNG), 0.6 ul of 20X TaqMan Gene Expression assay, and water to a final reaction volume of 12 ul. The amplification conditions consisted of an initial step at 95ºC for 10 min, followed by 55 cycles of 15 s at 95ºC and 1 min at 60ºC.

Table MM6. Primers sequences		
Primer pair	**Forward Primer**	**Reverse Primer**
CDH1	CGAGAGCTACACGTTCACGG	GGGTGTCGAGGGAAAAATAGG
FLG	TTTCGTGTTTGTCTGCTTGC	CTGGACACTCAGGTTCCCAT
KRT1	GTACCTGGTTCTGCTGCTCC	TGACCCTGAGATCCAAAAGG
KRT10	TGAGCCGCATTCTGAACGAG	GATGACTGCGATCCAGAGGA
IVN	ACTGAGGGCAGGGGAGAG	TCTGCCTCAGCCTTACTGTG
GAPDH	GGAGCGAGATCCCTCCAAAAT	GGCTGTTGTCATACTTCTCATGG

Immunoblot*:* protein lysates were prepared in RIPA buffer containing: 50 mM Tris-HCl pH 7.4, 1% NP-40, 0.2% sodium deoxycholate, 150 mM NaCl, 1 mM EDTA and 0.2% SDS. Proteins were detected by immunoblot following standard procedures.

Table MM7. Summary of the main materials and methods of Study II

Patients	**Extreme phenotype sampling:** grade 0 (controls) *v* grade 3 (cases) patients **Discovery cohort:** 166 capecitabine-treated patients (78 grade 0 *v* 88 grade 3) **Replication cohort:** 85 capecitabine-treated patients (61 grade 0 *v* 24 grade 3)
Genotyping	**Genome-wide association study (GWAS):** analysis of 616,795 SNVs across the human genome on the Illumina Human610 array
Statistical analysis	**Single-variant associations:** Cox proportional hazards regression, modeling the cumulative dose of capecitabine (mg/m^2) up to the development of grade 3 CiHFS for cases, or the total cumulative dose received in the case of controls.
Functional studies	• 4C-seq experiments in normal human epidermal keratinocytes • R-cadherin and involucrin protein expression in human keratinocytes and in skin samples from cancer patients before capecitabine treatment • R-cadherin expression knock-down

3. Materials & Methods, Study III and Study IV: identification of genetic variants predictive of susceptibility to chronic anthracycline-induced cardiotoxicity (AIC) in pediatric oncology and breast cancer patients

3.1. Patients

3.1.1. Pediatric oncology patients

93 anthracycline-treated pediatric cancer patients aged less than 30 years and treated at the *La Paz* University Hospital and the *Niño Jesús* University Hospital in Madrid and at the University Clinic of Navarra in Pamplona were reviewed between 2010 and 2014. All patients were treated with doxorubicin, daunorubicin or epirubicin as part of their chemotherapy protocol.

3.1.2. Breast cancer patients

Spanish breast cancer patients: 71 locally advanced adult breast cancer patients treated at the *San Carlos* University Hospital, (Madrid, Spain) were included. These patients were enrolled in a neoadjuvant phase II randomized clinical trial (www.clinicaltrials.gov, identifier code: NCT00123929). The eligibility patient criteria included the following (i) women aged between 18 and 79 years; (ii) clinical stage IIB, IIIA or IIIB breast cancer and palpable breast tumors not amenable to breast-preserving surgery. Patients were randomly assigned to receive four cycles of either neoadjuvant doxorubicin (75 mg/m^2) (39 patients) or neoadjuvant docetaxel (100 mg/m^2) (32 patients) every 3 weeks. After surgery, patient treatment assignment was crossed-over to receive four cycles of the opposite drug.

Belgium breast cancer patients: 142 early adult breast cancer patients treated at the Leuven Multidisciplinary Breast Cancer Center (University Hospitals Leuven, Belgium) were included. These patients were treated with 3–6 cycles of (neo) adjuvant 5-fluorouracil (500 mg/m^2), epirubicin (100 mg/m^2) and cyclophosphamide (500 mg/m^2).

All patients (pediatric and adult breast cancer patients) received anthracyclines as part of their chemotherapy protocol, had normal cardiac function before anthracycline chemotherapy and had echocardiographic evaluations (pre-chemotherapy and post-chemotherapy). Patients were excluded if they had a personal history of cardiac disease or were treated with concomitant (neo) adjuvant use of trastuzumab, because of its well-known association with cardiotoxicity. Written informed consent was obtained from adult patients and from the parents or legal guardians of children. The study was approved by the ethics committees of each participating hospital and university.

Patient medical records were reviewed retrospectively by oncologists and cardiologists. Demographic, clinical and therapeutic information extracted from medical records included demographics, disease characteristics, chemotherapy, diagnostic echocardiograms to document baseline and follow-up cardiac function and any cardiac compromise and its severity, and any symptoms or signs consistent with chronic AIC.

3.1.3. AIC definition

AIC was defined as early-onset (occurring within 1 year after anthracycline treatment completion) or late-onset (occurring >1 year after anthracycline therapy completion) chronic LV dysfunction in both, pediatric and adult breast cancer patients. To rule out acute AIC, only echocardiograms obtained 30 days or more after an anthracycline dose were considered.

AIC definition in pediatric oncology patients: AIC in pediatric patients was defined as chronic LV dysfunction assessed by echocardiogram measurements and evidenced by shortening fraction (SF) ≤27% at any time after anthracycline treatment completion (asymptomatic AIC) or symptoms/signs of cardiac complications, including severe mitral valve insufficiency, pericardial effusion, left ventricular hypertrophy or pulmonary hypertension (symptomatic AIC). The criteria for determining a symptomatic event were established by pediatric cardiologists. Controls were patients who had no symptoms or signs of cardiac complications and had normal echocardiograms (SF≥35%) during and after anthracycline therapy.

AIC definition in adult breast cancer patients: AIC in breast cancer patients was defined as early or late-onset (i) cardiac failure grade 3–5 using the NCI CTCAE v4.0 scoring system (grade 3: severe symptoms at rest or with minimal activity or exertion, intervention indicated; grade 4: life-threatening consequences, urgent intervention indicated; 5: death)[86] (ii) asymptomatic decrease of left ventricular ejection fraction (LVEF) ≥10%. Control patients were defined as those having no symptoms or signs of cardiac complications and normal echocardiograms (with a LVEF>60% at both baseline and follow-up and with a decline in LVEF≤5%) during and after anthracycline therapy.

3.2. Isolation and quantification of DNA

Genomic DNA was isolated from peripheral blood lymphocytes and quantified using the same procedures above described (section 1.3).

3.3. Genotyping

DNA samples from the Spanish pediatric and the adult breast cancer cohorts were genotyped on the Illumina HumanExome-12v1_A Beadchip (Illumina, San Diego, USA) array according to the manufacturers´ recommended protocols. The HumanExome arrays are based on the Illumina Infinium technology detailed for the genome-wide genotyping in **Study II** (section 2.3 and **Figure MM2**). As it was mentioned in the introduction, the HumanExome array includes 247,870 variants focused on protein-altering variants, selected from >12,000 exome and genome sequences representing multiple ethnicities and complex traits. Additional array content includes variants associated with complex traits in previous GWAS, HLA tags, ancestry informative markers, markers for identity-by-descent estimation, and random synonymous SNVs[47].

DNA samples from the Belgium adult breast cancer patients were genotyped for selected variants [rs79338777 (*ETFB*), rs149172980 (*WISP1*), rs72731540 (*WISP1*) and rs143089011 (*WISP1*)] using the MassARRAY genotyping system (Sequenom Inc., San Diego, USA) following the manufacturer's instructions (section 2.3 and **Figure MM5**).

3.4. Quality control (QC)

Genotype clustering and calling was carried out using GenTrain (v. 2.0) in GenomeStudio (v. 2011.1) (Illumina, San Diego, USA) in combination with zCall (v. 2.2)[163]. Anthracycline-treated pediatric and Spanish breast cancer samples were excluded from further analyses if they had more than 5% missing genotype data or showed a level of heterozygosity greater than 4.9 standard deviations from the mean. The remaining samples were subsequently assessed for population outliers and stratification using a PCA-based approach[155]. We carried out two PCAs, the first considering the Spanish pediatric cohort or the Spanish breast cancer cohort samples, to identify population outliers and the second including HapMap samples genotyped using the same array[164] using the snpStats package in R (v.3.0.1) (**Figure R21** and **Figure R24**). Variants were excluded if had a call rate<0.99 or deviated from Hardy-Weinberg equilibrium ($P<10^{-8}$).

3.5. Statistical analyses

We performed statistical analyses using PLINK (v. 1.07)[147], R (v.3.0.1) and SPSS (v.18.0; SPSS, Chicago, USA) software.

Covariates identification: associations between clinical factors and chronic AIC were assessed using Fisher´s exact test for categorical variables and Wilcoxon-Mann-Whitney U test for continuous variables. A *P*-value of <0.05 was considered to be statistically significant.

Single-variant associations: single-marker associations with risk of chronic AIC were performed assuming an additive genetic model and using logistic regression analysis. Covariates for the pediatric oncology cohort included age at diagnosis, cumulative anthracycline dose and bleomycin concomitant therapy. Covarites for the Spanish breast cancer cohort included age at diagnosis. Covariates for the overall series combined considering the pediatric oncology, the Spanish breast and the Belgium breast cancer cohorts together included age at diagnosis, cumulative anthracycline dose and whether patients had breast or childhood cancers. *P*-values were adjusted for multiple testing using Benjamini-Hochberg's false discovery rate (FDR-BH) procedure[165]. A *P*-value <0.05 was considered to be statistically significant after multiple testing correction.

Gene-based associations: for gene-based testing, we used the optimized sequence kernel association test (SKAT-O)[55,166] in GenABEL[167] with default weights[55] to assess the joint effects of common and low-frequency variants within each gene. Only genes with at least 3 genotyped variants were considered. Covariates for gene-based testing in the Spanish pediatric oncology cohort included age at diagnosis, cumulative anthracycline dose and bleomycin concomitant therapy. Covarites for gene-based testing in the Spanish breast cancer cohort included age at diagnosis. To adjust for multiple comparisons, the FDR-BH was applied[165]. A *P*-value of <0.05 was considered to be statistically significant after multiple testing correction. While the SKAT-O does not provide any parameter estimates, sensitivity analyses were applied to genes found to be associated, whereby we removed one variant at a time and repeated this SKAT-O test.

Likelihood ratio tests were used to select the best fitting model for the association with chronic AIC and to identify independent association signals across multiple variants, comparing the model considering both each genetic variant and clinical factors with the model with only clinical factors.

3.6. Gene enrichment analysis

To gain an in-depth understanding of the biological interpretation of large gene lists that are enriched in exome array analyses, we used the functional annotation clustering analysis module

of the bioinformatic tool DAVID[168]. This type of grouping of functional annotations gives a more insightful view of the relationships between annotation categories and terms than the traditional linear list of enriched terms[169]. Each annotation term group is assigned an enrichment score (ES) to rank overall importance. ES≥1.3 indicates biological significance. DAVID also provides a *P*-value to assess the significance of gene-term enrichment, which is corrected by applying the FDR-BH procedure[165].

3.7. Pathway enrichment analysis

We also carried out an exploratory SKAT-O analysis on a combined pathway set consisting of 1912 gene sets taken from the Gene Ontology[170,171], Kyoto Encyclopedia of Genes and Genomes (KEGG)[172,173], and Reactome[174,175] repositories.

3.8. GPR35 Sanger sequencing

Primers spanning the two exons of the *GPR35* transcript ENST00000430267 were used to amplify germline DNA from the 93 anthracycline-treated pediatric cancer patients. A 244 bp fragment containing the exon 1 of the *GPR35* transcript ENST00000430267 was amplified starting from 100 ng of genomic DNA using the following primers 5'-ACAAATACATTCTGGAGATGACC-3' and 5'-TGACCCAATAACCCCACTTC-3' (amplification product=244 bp). Exon 2 was amplified using 3 pairs of primers: i) first pair: 5'-CAGAGGTGGGCAGAGTGG-3' and 5'-CAGCCTGCCTGGGGGAC-3' (amplification product=531 bp); ii) second pair: 5'-ACTCCCTGCGAGACACCTC-3' and 5'-TTCAGGGAGCAGAAGACCAC-3' (amplification product=346 bp); iii) third pair: 5'-ATTCTACCTGCCCCTGGCC-3' and 5'-TGCTACTGGTTCCAGCTTCC-3' (amplification product=476 bp); starting from 100 ng of genomic DNA. The purified products were subsequently sequenced using the automatic sequencer ABI 3730xl (Applied Biosystems).

3.9. In silico prediction

We used the SIFT algorithm[176] and the programs PolyPhen-2[177], MutPred[178], SNPs&GO[179] and PON-P2[180] to predict the impact of selected variants on protein structure or function. SIFT is a sequence homology-based tool to predict tolerated and deleterious substitutions for every position of a given query sequence. PolyPhen-2 classifies variants as "benign," "possibly damaging" or "probably damaging" based on a sequence-based characterization of the substitution site, profile analysis of homologous sequences, as well as their location in the three-dimensional structure of the protein molecule. MutPred not only uses multiple alignment

information based upon SIFT but also includes information on the gain or loss of 14 different structural and functional properties to assign a pathogenicity score; we considered scores greater than 0.5 to define pathogenicity. SNPs&GO predicts whether a variant is disease-related or neutral by exploiting protein sequence, evolutionary information, and function. PON-P2 utilizes information about evolutionary conservation of sequences, physical and biochemical properties of amino acids, GO annotations and if available, functional annotations of variation sites, to classify amino acid substitutions in human proteins. In addition, we used PredictSNP[181] to obtain a consensus prediction of pathogenicity based on MAPP, PhD-SNP, PolyPhen-1, Polyphen-2, SIFT, SNAP nsSNPAnalyzer, and PANTHER. We also used the F-SNP database[182] which extracts information from large number of resources, such as PolyPhen, SIFT, SNPeffect, SNPs3D, LS-SNP, Ensembl, ESEfinder, RescueESE, ESRSearch, PESX, TFSearch, Consite, GoldenPath, KinasePhos, OGPET, and Sulfinator, to predict functional effects at the splicing, transcriptional, translational, and post-translational level.

Table MM8. Summary of the main materials and methods of Study III

Patients	93 Spanish anthracycline-treated pediatric cancer patients (58 controls *v* 35 cases)
Patient definition	**Symptomatic case:** anthracycline-treated patients with symptoms/signs of cardiac complications, including severe mitral valve insufficiency, pericardial effusion, left ventricular hypertrophy or pulmonary hypertension **Asymptomatic case:** anthracycline-treated patients with LV dysfunction evidenced by shortening fraction (SF)≤27% at any time after anthracycline treatment completion **Control:** anthracycline-treated patients with no symptoms or signs of cardiac complications and with normal echocardiograms (SF≥35%) during and after anthracycline therapy
Genotyping	**Exome array:** analysis of 247,870 coding variants on the Illumina Human Exome Beadchip, which is enriched for low-frequency and rare coding variants (>80% variants with MAF ≤1%)
Statistical analysis	**Single-variant associations:** logistic regression analysis considering and additive genetic model. Significance level: *P*-value of <0.05 after multiple testing correction. **Gene-based associations:** analysis of the joint effect of variants within each gene (≥3 variants) using SKAT-O. Significance level: *P*-value of <0.05 after multiple testing correction. **Sensitivity analyses:** the individual contribution of variants within *GPR35* was assessed by removing one variant at a time and recalculating the association for *GPR35* using SKAT-O.
Functional annotations	*In silico* prediction: SIFT, PolyPhen-2, MutPred, SNPs&GO, PON-P2, PredictSNP and F-SNP.

Table MM9. Summary of the main materials and methods of Study IV

Patients	**Spanish breast cancer cohort:** 71 anthracycline-treated advanced breast cancer patients (53 controls *v* 18 cases) **Belgium breast cancer cohort:** 142 anthracycline-treated early breast cancer patients (86 controls *v* 56 cases) **Spanish pediatric cancer patients:** 93 anthracycline-treated pediatric oncology patients (58 controls *v* 35 cases)
Patient definition	**Breast cancer patients** **Symptomatic case:** anthracycline-treated patients with symptomatic cardiac failure grade 3–5 (CTCAE v.4.0) **Asymptomatic case:** anthracycline-treated patients with LV dysfunction evidenced by an asymptomatic decrease of left ventricular ejection fraction (LVEF) ≥10%. **Control:** anthracycline-treated patients with no symptoms or signs of cardiac complications and with normal echocardiograms (LVEF>60% at both baseline and follow-up and with a decline in LVEF≤5%) during and after anthracycline therapy. **Spanish pediatric cancer patients** see **Table MM8**
Genotyping	**Exome array:** analysis of 247,870 coding variants on the Illumina Human Exome Beadchip, which is enriched for low-frequency and rare coding variants (>80% variants with MAF ≤1%)
Statistical analysis	**Single-variant associations:** logistic regression analysis considering and additive genetic model. Significance level: *P*-value of <0.05 after multiple testing correction. **Gene-based associations:** analysis of the joint effect of variants within each gene (≥3 variants) using SKAT-O. Significance level: *P*-value of <0.05 after multiple testing correction. **Sensitivity analyses:** the individual contribution of variants within *ETFB* and *WISP1* genes was assessed by removing one variant at a time and recalculating the association for the gene using SKAT-O.
Functional annotations	*In silico* prediction: SIFT, PolyPhen-2, MutPred, SNPs&GO, PON-P2, PredictSNP and F-SNP.

RESULTS

1. Results, Study I: identification of predictive and prognostic genetic variants for Ewing sarcoma

The demographic and clinical characteristics of both cohorts are shown in **Table R1**.

Table R1. Clinical characteristics of Ewing sarcoma patients

	Discovery (N=106)		Replication (N=389)	
Characteristic	**N**	**%***	**N**	**%***
Age at diagnosis (years)				
Median	12.2		14.5	
Range	0.4-27.8		0.1-27.8	
Sex				
Female	42	39.6	156	40.1
Male	64	60.4	233	59.9
Primary site				
Upper extremities	9	8.7	39	10.2
Lower extremities	48	46.2	129	33.9
Axial	41	39.4	203	53.3
Soft tissue	6	5.8	10	2.6
Missing	2		8	
Metastasis at diagnosis				
No	62	59.6	255	67.3
Yes	42	40.4	124	32.7
Missing	2		10	
Response to treatment				
Good	55	70.5	148	64.3
Poor	23	29.5	82	35.7
Missing	28		159	
Relapse				
No	67	67.0	211	58.1
Yes	33	33.0	152	41.9
Missing	6		26	
Vital status				
Alive	67	67.0	229	61.7
Dead	33	33.0	142	38.3
Missing	6		18	
Follow-up (years)				
Median	93.7		70.3	
Range	7.8-300		12.3-312	

*Percentages are computed based on the total number of non-missing values

The median age at diagnosis was 12.2 years in the discovery cohort and 14.5 years in the replication series. Consistent with previous reports[61], Ewing sarcoma was slightly more common in men than in women. Axial and lower extremity primary sites were the most common locations of primary tumors, while Ewing sarcoma affecting soft tissue was only found in 16 patients. At the time of diagnosis, 42 (40.4%) of the 104 evaluable Spanish patients and 124 (32.7%) of the 379 evaluable European patients already presented metastases, while 33 (33%) and 152 (41.9%) patients, respectively, developed metastasis or local recurrences during follow-

up. More than 60% of patients in both cohorts were good responders and were still alive at last follow-up, with a median follow-up after diagnosis of 93.7 months for the Spanish patients and 70.3 months in the replication cohort.

After filtering, 334 SNVs of the 384 genotyped variants were successfully analyzed (**Figure R1**). There was no evidence of departure from Hardy–Weinberg equilibrium for any of them. Data for two patients in the Spanish cohort and 55 patients in the replication cohort were excluded due to a low genotyping call rate (<0.90), leaving 104 and 334 patients, respectively.

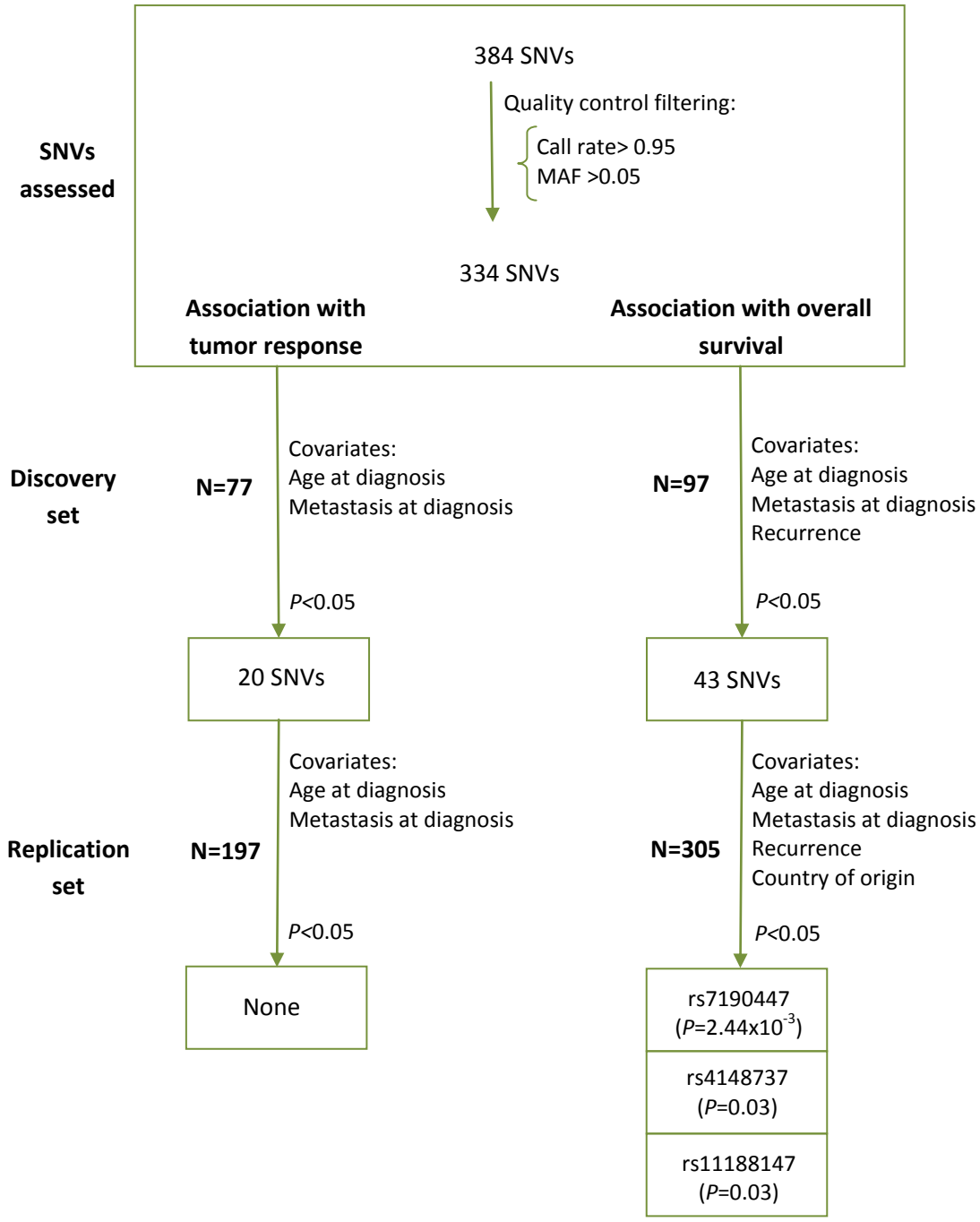


Figure R1. Flow chart of Study I.

1.1. Associations with tumor response to treatment

Associations with tumor response were assessed in 77 patients from the Spanish discovery cohort for which this information was available. After adjusting for age and the presence of metastasis at diagnosis, an association with $P<0.05$ was observed for 20 SNVs. However, none of these associations were replicated in the European cohort (N=197, $P\geq0.05$) (**Figure R1**).

Also including neoadjuvant therapy as an additional covariate in multivariable models made no substantial difference to the results obtained.

1.2. Associations with overall survival

Since adjustment for tumor response made no substantial difference to the estimated HR (based on an analysis of the cases for which this information was available), we present results without adjustment for this covariate, based on a larger sample size.

We identified 43 SNVs associated with overall survival at $P<0.05$ in the Spanish cohort after adjusting for age at diagnosis, presence of metastasis at diagnosis, and recurrence (N=97) (**Table R2**). Associations with three of these were replicated in the European cohort (N=305) (**Table R2**, **Table R3**). The strongest evidence of association was found for the SNV rs7190447, an intronic polymorphism in *ABCC6* (ATP-binding cassette sub-family C member 6). In both cohorts, a recessive model was the best fit; C-allele homozygotes had a higher risk of death [discovery phase: HR=14.30, 95%CI=1.53-134, $P=0.020$; replication phase: HR=9.28, 95%CI=2.20-39.2, $P=0.0024$, **Figure R2A**]. The 5-year survival for patients carrying 1 or 2 copies of the G allele was 74% in the discovery cohort and 65% in the replication series, while no individual carrying the CC genotype lived for 5 years following diagnosis (**Table R3**).

The G allele of rs4148737, an intronic variant in *ABCB1* (ATP-binding cassette sub-family B member 1), was associated with poorer overall survival under a recessive model [discovery phase: HR=2.96, 95%CI=1.08-8.10, $P=0.034$; replication phase: HR=1.60, 95%CI=1.05-2.44, $P=0.029$, **Table R2**, **Table R3**, **Figure R2B**]. The estimated 5-year survival for cases with AA/AG and GG genotypes was 75% and 66%, respectively, in the discovery cohort, and 66% and 55%, respectively, in the European cohort (**Table R3**).

Finally, the minor T allele for a SNV located 2.7kb downstream of the *CYP2C8* (cytochrome P450 family 2 subfamily C member 8) gene, rs11188147, was associated with an increased risk of

death under a recessive model [discovery phase: HR=2.49, 95%CI=1.06-5.87, *P*=0.037; replication phase: HR=1.77, 95%CI=1.06-2.96, P=0.030, **Table R2**, **Table R3**, **Figure R2C**]. The 5-year survival for CC/CT and TT carriers was 76% and 67%, respectively, in the discovery cohort, and 65% and 58%, respectively in the replication series (**Table R3**).

In order to evaluate the possible functional significance of the replicated and strongly correlated variants we applied the HaploReg web tool and considered ENCODE data. Functional annotations of SNVs rs7190447, rs4148737 and rs11188147 and correlated variants ($r^2 \geq 0.8$) are shown in **Table R4** and **Figure R3**, **Figure R4** and **Figure R5**.

SNV rs7190447 (*ABCC6*) was found to be in perfect LD with rs7192303, an intronic polymorphism located 122 bp upstream. The genomic region containing rs7192303 is enriched with specific histone marks associated with transcribed regions in a hepatocellular carcinoma cell line, with weak enhancers in a skeletal muscle myoblast cell line, and with a DNase hypersensitive cluster in 123 different cell types (**Table R4** and **Figure R3**). The strongest and most robust chromatin immunoprecipitation (ChIP)-seq signal is observed for CCCTC-binding factor (CTCF) binding in a large number of ENCODE cell lines (69) (**Table R4** and **Figure R3**). Remarkably, strong signals for cohesin subunits Rad21 and SMC3, were also observed. The genomic region containing rs7192303 also has the potential to form chromatin loops, through CTCF binding, with intronic regions of *ABCC6* and *ABCC*1, both located upstream, in a breast cancer cell line (MCF-7) (**Figure R4**). We explored expression quantitative trait loci (eQTLs) using Genotype-Tissue Expression (GTEx) data, and interestingly, we found statistically significant differences in *ABCC6* gene expression by rs7192303 genotype in the liver (*P*=0.019, effect size=0.23). These findings suggest rs7192303 as the most plausible causal variant for the observed association with overall survival.

The intronic variant rs4148737 (*ABCB1*) resides in a weakly transcribed region in a hepatocellular carcinoma cell line, but also overlaps with a weak enhancer in GM12878 and in a RUNX3 ChIP-seq cluster in the same lymphoblastoid cell line. It was predicted to overlap with a DNAse hypersensitive region in a lymphoblastoid cell line and in cerebellar and hippocampal astrocytic cell lines, and to alter EBF, ERalpha-a, Hic1 regulatory motifs (**Table R4** and **Figure R5**). None of the nine intronic SNVs in high LD ($r^2 \geq 0.84$) with this variant had stronger functional evidence reported (**Table R4**). According to GTEx data, rs4148737 influences *ABCB1* expression in testis (P=0.000016, effect size= -0.44).

No strong functional evidence was observed for SNV rs11188147 (*CYP2C8*) or for any of the 18 variants that are in high LD with it ($r^2 \geq 0.83$) (**Table R4**). No significant differences in *CYP2C8* expression were found for any tissues available in GTEx.

Main results Study I

We identified associations with overall survival (*P*<0.05) for three SNVs in the Spanish cohort that were replicated in the European cohort. The strongest association observed was with rs7190447, located in the ATP-binding cassette subfamily C member 6 (*ABCC6*) gene [discovery: hazard ratio (HR)=14.30, 95% confidence interval (CI)=1.53–134, *P*=0.020; replication: HR=9.28, 95%CI=2.20–39.2, *P*=0.0024] and its correlated variant rs7192303, which was predicted to have a plausible regulatory function. We also replicated associations with rs4148737 in the ATP-binding cassette subfamily B member 1 (*ABCB1*) gene (discovery: HR=2.96, 95% CI=1.08–8.10, P=0.034; replication: HR=1.60, 95%CI=1.05–2.44, *P*=0.029), and rs11188147 in cytochrome P450 family 2 subfamily C member 8 gene (*CYP2C8*) (discovery: HR=2.49, 95% CI=1.06–5.87, *P*=0.037; replication: HR=1.77, 95% CI=1.06–2.96, *P*=0.030). None of the associations with tumor response were replicated.

Table R2. Analysis of associations between variants and overall survival

						Discovery (N=97)			Replication (N=305)		
Gene	**Chr.**	**Variant**	**Position ***	**Location**	**Model**	***P***	**Minor allele HR**	**95%CI**	***P***	**Minor allele HR**	**95%CI**
ABCC1	16	rs212081	16225971	Intronic	Recessive	$2.59x10^{-4}$	T 6.62	2.40-18.3	0.46	T 1.19	0.75-1.90
SLCO6A1	5	rs981988	101819981	Intronic	Recessive	$1.52x10^{-3}$	G 54.2	4.59-639	0.11	G 0.20	0.03-1.41
GSTP1	11	rs7927381	67346743	4.3 kb upstream	Recessive	$1.52x10^{-3}$	T 54.2	4.59-639	0.28	T 3.06	0.39-23.8
ABCB1	7	rs7787082	87157051	Intronic	Additive	$1.91x10^{-3}$	A 0.29	0.13-0.64	0.90	A 1.00	0.71-1.42
ABCC3	17	rs12451302	48751164	Intronic	Recessive	$2.62x10^{-3}$	G 5.16	1.77-15.0	0.29	G 0.80	0.52-1.22
CYP2C9	10	rs4918758	96697252	2 kb upstream	Recessive	$2.98x10^{-3}$	C 5.11	1.74-15.0	0.09	C 1.55	0.93-2.59
ABCC3	17	rs3785912	48756937	Intronic	Recessive	$4.06x10^{-3}$	A 5.99	1.77-20.3	0.25	A 0.70	0.38-1.29
ABCC1	16	rs3888565	16183045	Intronic	Additive	$4.69x10^{-3}$	A 0.28	0.12-0.68	0.25	A 0.83	0.59-1.15
CYP3A5	7	rs28365067	99272310	Intronic	Additive	$5.84x10^{-3}$	T 4.42	1.54-12.7	0.89	T 1.04	0.62-1.74
ABCC1	16	rs35621	16168608	Intronic	Additive	$7.27x10^{-3}$	T 0.21	0.07-0.66	0.97	T 0.99	0.68-1.46
ABCB1	7	rs2214102	87229501	5´ UTR	Recessive	$8.15x10^{-3}$	A 28.4	2.38-339	-	-	-
ABCC1	16	rs16967755	16199255	Intronic	Additive	0.01	G 0.31	0.13-0.76	0.64	G 1.07	0.80-1.45
ABCC6	16	rs16967488	16252696	Intronic	Dominant	0.01	C 2.80	1.24-6.35	0.73	C 1.07	0.74-1.53
ALDH1A1	9	rs348481	75514436	1.1 kb downstream	Recessive	0.01	C 8.03	1.53-42.0	0.99	C 0.99	0.31-3.15
ABCC2	10	rs717620	101542578	5´ UTR	Recessive	0.02	A 15.8	1.72-146	0.91	A 0.93	0.23-3.77
ABCC1	16	rs35626	16170615	Intronic	Additive	0.02	T 0.41	0.20-0.85	0.90	T 0.98	0.74-1.30
CYP2C9	10	rs11597626	96604273	Intronic	Dominant	0.02	G 0.37	0.17-0.84	0.89	G 1.03	0.71-1.48
CYP2C9	10	rs12251688	96693727	4.6 kb upstream	Dominant	0.02	T 0.38	0.17-0.84	0.83	T 1.04	0.72-1.49
CYP3A4	7	rs4646437	99365083	Intronic	Dominant	0.02	T 3.06	1.20-7.79	0.43	T 1.19	0.77-1.83
ABCC6	**16**	**rs7190447**	**16289126**	**Intronic**	**Recessive**	**0.02**	**C 14.30**	**1.53-134**	**$2.44x10^{-3}$**	**C 9.28**	**2.20-39.2**

Continue on next page

Continued

Table R2. Analysis of associations between variants and overall survival (continued)

						Discovery (N=97)			Replication (N=305)		
Gene	**Chr.**	**Variant**	**Position ***	**Location**	**Model**	***P***	**Minor allele HR**	**95%CI**	***P***	**Minor allele HR**	**95%CI**
ABCC1	16	rs12922404	16060994	Intronic	Dominant	0.02	T 2.61	1.16-5.89	0.17	T 1.29	0.89-1.88
ABCB1	7	rs2235048	87138511	Intronic	Recessive	0.02	C 2.88	1.16-7.12	0.46	C 0.85	0.55-1.31
ABCB1	7	rs17064	87133470	3´ UTR	Additive	0.02	T 4.08	1.22-13.7	0.96	T 0.99	0.60-1.61
ABCC6	16	rs2238469	16283071	Intronic	Recessive	0.03	A 4.36	1.17-16.3	0.15	A 0.23	0.03-1.73
ABCC3	17	rs8079432	48749883	Intronic	Additive	0.03	G 2.77	1.11-6.91	0.68	G 0.91	0.57-1.45
ABCC4	13	rs9590220	95906694	Intronic	Additive	0.03	T 0.38	0.16-0.91	0.13	T 1.28	0.93-1.75
ABCB1	7	rs10264990	87202615	Intronic	Recessive	0.03	C 3.21	1.11-9.26	0.10	C 1.54	0.92-2.59
GSTP1	11	rs614080	67347287	3.8 kb upstream	Recessive	0.03	G 2.56	1.09-6.04	0.52	G 0.85	0.52-1.39
ABCC1	16	rs212087	16230290	Intronic	Recessive	0.03	T 0.34	0.13-0.91	0.10	T 1.46	0.93-2.30
ABCB1	**7**	**rs4148737**	**87171152**	**Intronic**	**Recessive**	**0.03**	**G 2.96**	**1.08-8.10**	**0.03**	**G 1.60**	**1.05-2.44**
ABCG2	4	rs2725264	89026109	Intronic	Additive	0.04	G 2.60	1.06-6.36	0.41	G 0.83	0.53-1.29
CYP1B1	2	rs4646429	38306935	3.6 kb upstream	Additive	0.04	A 0.44	0.20-0.95	0.95	A 0.99	0.73-1.34
CYP2C8	**10**	**rs11188147**	**96793820**	**2.7 kb downstream**	**Recessive**	**0.04**	**T 2.49**	**1.06-5.87**	**0.03**	**T 1.77**	**1.06-2.96**
MPO	17	rs7208693	56357818	Missense	Additive	0.04	A 2.17	1.04-4.51	0.76	A 0.93	0.54-1.47
ABCC1	16	rs4148354	16174506	Intronic	Dominant	0.04	G 0.42	0.19-0.96	0.87	G 1.07	0.69-1.57
CYP2C8	10	rs1934956	96828160	Intronic	Recessive	0.04	T 4.95	1.06-23.2	0.47	T 1.44	0.53-3.96
CYP2A6	19	rs8192729	41350996	Intronic	Additive	0.05	A 2.99	1.02-8.75	0.92	A 1.03	0.62-1.69
ABCC1	16	rs2299670	16220858	Intronic	Additive	0.05	G 0.51	0.26-0.99	0.69	G 1.06	0.79-1.43
SOD1	21	rs2070424	33039320	Intronic	Additive	0.05	G 3.22	1.02-10.2	0.06	G 1.54	0.98-2.41
LPO	17	rs8178407	56344656	Intronic	Dominant	0.05	G 2.61	1.01-6.77	0.77	G 0.95	0.66-1.37
ABCC1	16	rs11075295	16177687	Intronic	Recessive	0.05	G 9.14	1.02-82.1	0.96	G 0.98	0.36-2.65
ABCB1	7	rs13237132	87191669	Intronic	Additive	0.05	G 1.80	1.00-3.22	0.20	G 1.19	0.91-1.54
ABCC4	13	rs9590211	95892414	Intronic	Dominant	0.05	A 2.14	1.00-4.57	0.71	A 0.93	0.63-1.37

HRs are per copy of the specified minor allele. Variants shown in bold were those associated with overall survival in both cohorts at $P<0.05$. * Chromosome positions are based on Genome Reference Consortium Human Build 37 (GRCh37/hg19). Abbreviations: Chr, chromosome; kb, kilobases; HR, hazard ratio; CI, confidence interval.

Table R3. Associations between variants and overall survival in Ewing sarcoma patients

							Discovery (N=97)						Replication (N=305)				
Gene	**Chr.**	**Variant**	**Position ***	**Location**	**Model**	**MAF**	**Genotype**	**N**	**5-year OS**	***P***	**HR**	**95%CI**	**N**	**5-year OS**	***P***	**HR**	**95%CI**
ABCC6	16	rs7190447	16289126	Intronic	Recessive	0.07	GG/GC	96	74%				303	65%			
							CC	1	-				2	-			
							per allele C			0.020	14.3	1.53-134			0.0024	9.28	2.20-39.2
ABCB1	7	rs4148737	87171152	Intronic	Recessive	0.43	AA/AG	80	75%				251	66%			
							GG	17	66%				54	55%			
							per allele G			0.034	2.96	1.08-8.10			0.029	1.60	1.05-2.44
CYP2C8	10	rs11188147	96793820	2.7 kb downstream	Recessive	0.39	CC/CT	72	76%				263	65%			
							TT	25	67%				42	58%			
							per allele T			0.037	2.49	1.06-5.87			0.030	1.77	1.06-2.96

HRs are per copy of the specified minor allele. * Chromosome positions are based on Genome Reference Consortium Human Build 37 (GRCh37/hg19). Abbreviations: Chr., chromosome; kb, kilobases; 5-year OS, 5-year overall survival; HR, hazard ratio; CI, confidence interval

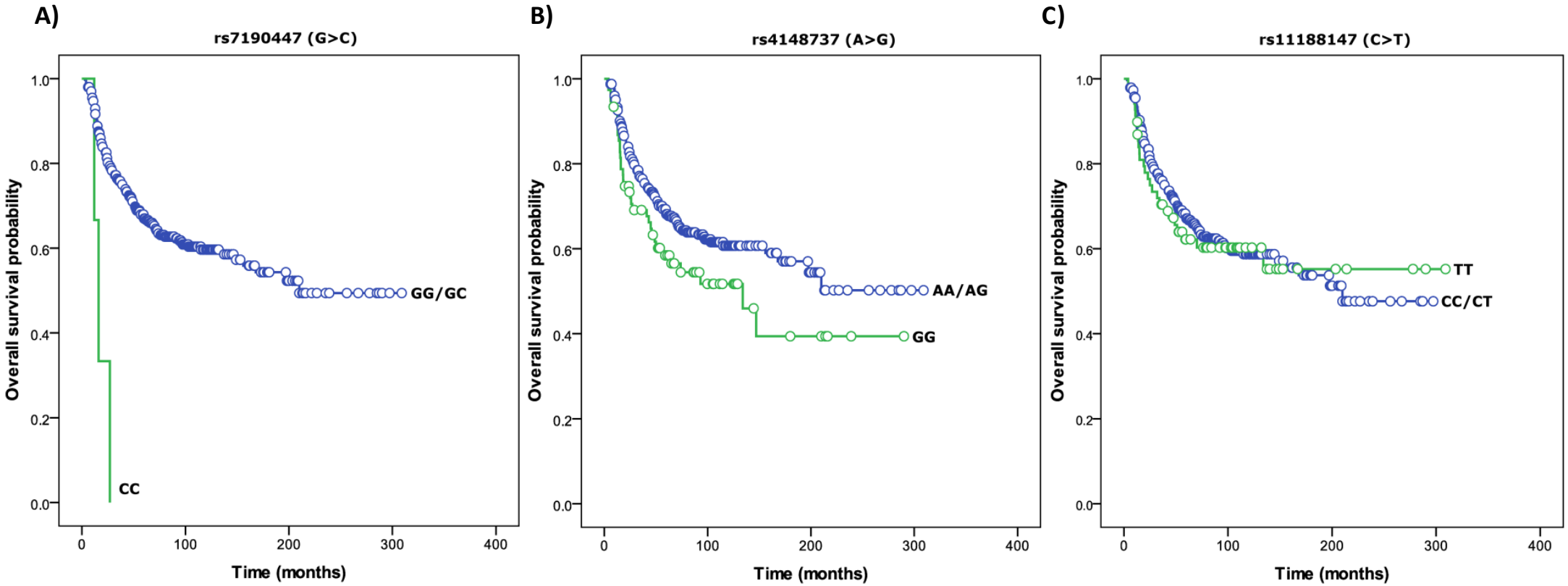


Figure R2. Kaplan-Meier survival curves for Ewing sarcoma patients (discovery and replication cohorts combined) according to genotype for (**A**) rs7190447 in *ABCC6* ($N_{GG/GC}$=399, N_{CC}=3, χ^2=14.84, $P_{log\text{-}rank}$=1.17×10^{-4}); (**B**) rs4148737 in *ABCB1* ($N_{AA/AG}$=326, N_{GG}=76, χ^2=3.73, $P_{log\text{-}rank}$=0.053); and (**C**) rs11188147 in *CYP2C8* ($N_{CC/CT}$=333, N_{TT}=69, χ^2=0.15, $P_{log\text{-}rank}$=0.69).

Table R4. Analysis of functional annotations of rs7190447, rs4148737 and rs11188147 and correlated variants ($r^2 \geq 0.8$)

Variant	Correlated variant	LD (r^2)	Enhancer histone marks	DNase	Proteins bound	Motifs changed
rs7190447	-	-	-	HA-sp	CTCF	GATA, Pou3f2
rs7206048	rs7190447	0.96	K562	GM12892	MAFK	Egr-1,GATA
rs6498619	rs7190447	1	K562	-	-	Pax-5
rs8044613	rs7190447	1	-	-	-	Evi-1,Pou2f2
rs11862259	rs7190447	1	-	-	-	Mef2, NF-AT1
rs7184822	rs7190447	0.98	-	-	-	22 altered motifs
rs7186376	rs7190447	1	-	-	-	ERalpha-a, HNF4, TLX1::NFIC
rs7187235	rs7190447	1	-	-	-	AP-1
rs7186601	rs7190447	1	-	-	-	BDP1_disc2, GCNF, Nr2f2, p300_known1
rs7192303	rs7190447	1	HSMM	PanIsletD, AG09309, AG10803, HA-h, HA-sp, HGF, HIPEpiC, HNPCEpiC, HPdLF, HVMF	CTCF, SMC3, ZNF143, RAD21, FOXA1,FOXA2, GATA3	INSM1
rs7199104	rs7190447	1	-	-	-	Foxa,Foxj2, Osf2
rs4148737	-	-	GM12878	GM12865,HA-h,HAc	-	EBF, ERalpha-a, Hic1
rs35572298	rs4148737	0.84	-	-	-	8 altered motifs
rs35280822	rs4148737	0.89	-	-	-	Homez,Lhx3
rs12154941	rs4148737	0.91	-	-	-	AP-1,Zfx
rs4148736	rs4148737	1	-	-	-	GR,Nanog
rs6961419	rs4148737	1	-	HConF,HFF-Myc,NHDF-neo	-	-
rs6961882	rs4148737	1	-	-	-	6 altered motifs
rs4148735	rs4148737	1	GM12878	Melano	-	GR, p300
rs1922242	rs4148737	1	-	-	-	5 altered motifs

Continue on next page

Table R4. Analysis of functional annotations of rs7190447, rs4148737 and rs11188147 and correlated variants (r^2≥0.8) (continued)

Variant	Correlated variant	LD (r^2)	Enhancer histone marks	DNase	Proteins bound	Motifs changed
rs2091766	rs4148737	0.88	-	-	-	8 altered motifs
rs11188147	-	-	-	-	-	-
rs1578436	rs11188147	1	-	-	-	9 altered motifs
rs7073968	rs11188147	1	-	-	-	HNF4, NF-I
rs10882517	rs11188147	1	-	-	-	Foxa, GZF1, Pou1f1
rs11188149	rs11188147	0.99	-	-	-	Pax-4
rs947173	rs11188147	0.99	-	-	-	BCL, BDP1, NRSF
rs1891070	rs11188147	0.99	-	-	-	Foxp1, Hdx
rs11572133	rs11188147	0.99	-	-	-	NRSF, Sin3Ak-20
rs12773510	rs11188147	0.99	-	-	-	11 altered motifs
rs199539470	rs11188147	0.95	-	-	-	10 altered motifs
rs58385086	rs11188147	0.99	-	-	-	11 altered motifs
rs11188156	rs11188147	0.98	-	-	-	DMRT3,Gfi1b
rs10882521	rs11188147	0.93	-	-	-	4 altered motifs
rs9702453	rs11188147	0.83	-	-	-	HNF4, Pdx1
rs145809484	rs11188147	0.90	-	-	-	GCM, Gcm1
rs143042734	rs11188147	0.88	-	-	-	5 altered motifs
rs13313110	rs11188147	0.98	-	-	-	Gfi1,Hsf,TATA
rs3752988	rs11188147	0.99	-	-	-	Ik-2,Mef2
rs10882525	rs11188147	0.99	-	-	-	4 altered motifs

The replicated variants associated with overall survival (in bold) and SNVs in strong (r^2>0.8) linkage disequilibrium (LD) with the replicated variants were analyzed using HaploReg to explore if they affected chromatin states or altered regulatory motifs or binding sites. AG09309, adult toe fibroblast cells; AG10803, abdominal skin fibroblast cells; GM12865, lymphoblastoid cells; GM12878, lymphoblastoid cells; GM12892, lymphoblastoid cells; HAc, human cerebellar astrocytic cells; HA-h, human hippocampal astrocytic cells; HA-sp, human spinal cord astrocytic cells; HConF, conjunctival fibroblast cells; HFF-Myc, foreskin fibroblast cells expressing canine cMyc; HGF, gingival fibroblasts cells; HIPEpiC, iris pigment epithelial cells; HNPCEpiC, non-pigment ciliary epithelial cells; HPdLF, periodontal ligament fibroblasts cells; HSMM, skeletal muscle myoblast cells; HVMF, villous mesenchymal fibroblast cells; K562, erythrocytic leukemia cells; Melano, human epidermal melanocyte cells; NHDF-neo,neonatal dermal fibroblast cells; PanIsletD, dedifferentiated human pancreatic islets

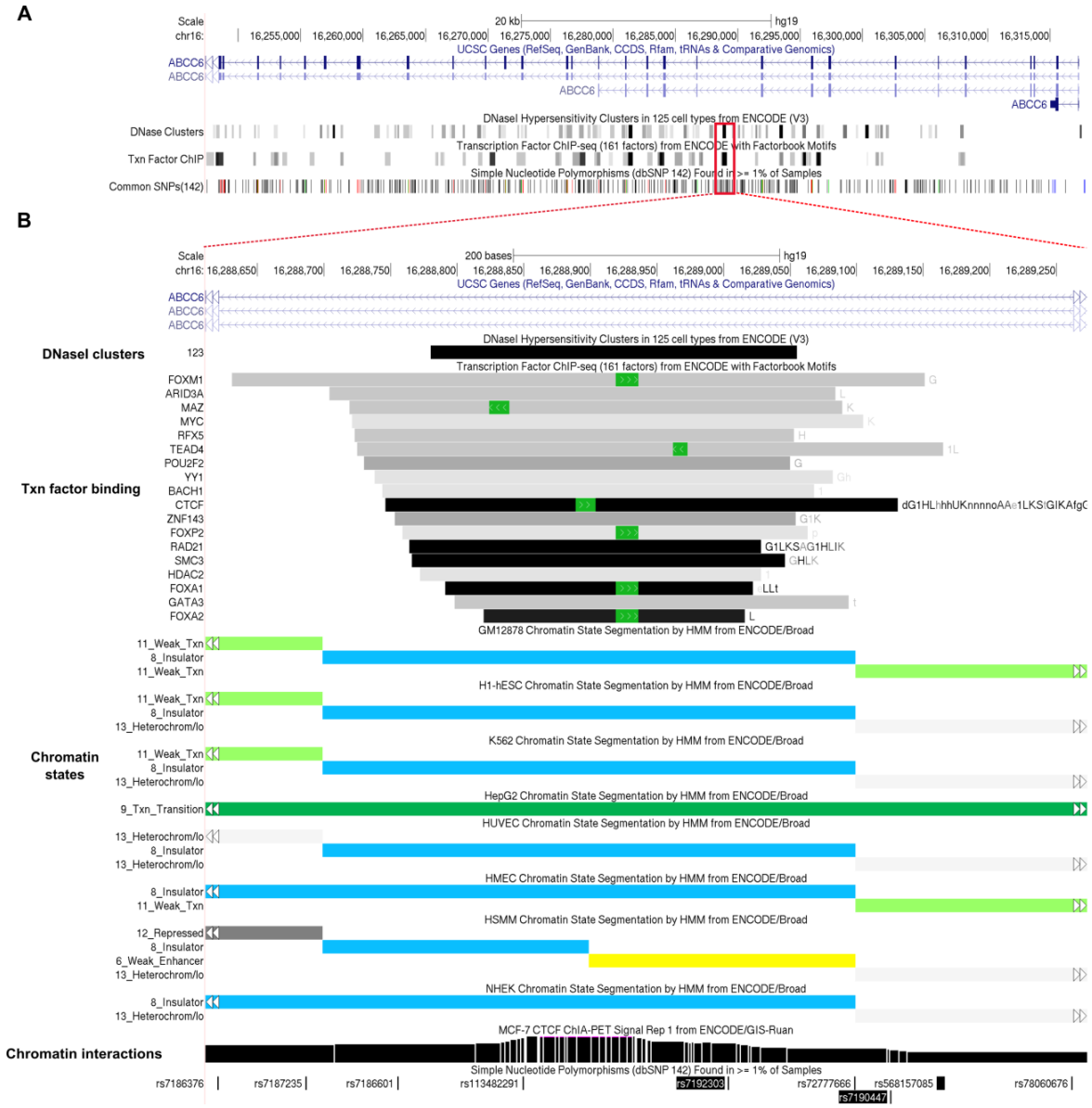


Figure R3. ENCODE functional evidence displayed in the UCSC Genome Browser for rs7190447 and nearby SNVs. (**A**) Genomic location of the *ABCC6* gene. Multiple DNase-seq and transcription factor ChIP-seq clusters can be observed. (**B**) Genomic location of rs7190447 (highlighted). A DNase hypersensitivity region was observed in 123 ENCODE cells around rs7190447 and overlapping with rs7192303 (highlighted), an intronic polymorphism located 122 bp upstream from rs7190447 and in perfect LD. Multiple transcription factors binding clusters can be observed in a large number of cells (identified by single-letter abbreviations). Gray boxes indicate the extent of the hypersensitive region for DNaseI hypersensitivity clusters or transcription factor occupancy, with the darkness of the box proportional to the maximum signal strength observed in any cells contributing to the cluster. The number to the left of a DNaseI hypersensitivity box shows how many cells are hypersensitive in the region. Within a ChIP-seq cluster, green highlighting indicates the highest scoring site of a Factorbook-identified canonical motif for the corresponding factor. Chromatin states characterized by combinations of histone marks are also shown in different human cell lines. Each chromatin state is associated with a different segment color. Blue, insulator; light green, weak transcribed; dark green, transcriptional transition; yellow, weak enhancer. The genomic region containing rs7190447 and rs7192303 is also enriched for CTCF-mediated chromatin interactions in MCF-7 breast cancer cells. MCF-7 CTCF ChIA-PET interactions are shown as a density graph of signal enrichment based on aligned read density. Due to space limitations, only a subset of cells where a CTCF-ChIP-seq peak is detected and CTCF ChIA-PET interactions are shown. GM12878, lymphoblastoid cells; K1-hESC, embryonic stem cells; K562, erythrocytic leukemia cells; HepG2, hepatocellular carcinoma cells; HUVEC, umbilical vein endothelial cells; HMEC, mammary epithelial cells; HSMM, skeletal muscle muscle myoblast cells; NHEK, normal epidermal keratinocytes cells. Abbreviations: SNP, single nucleotide polymorphism; Txn, transcription.

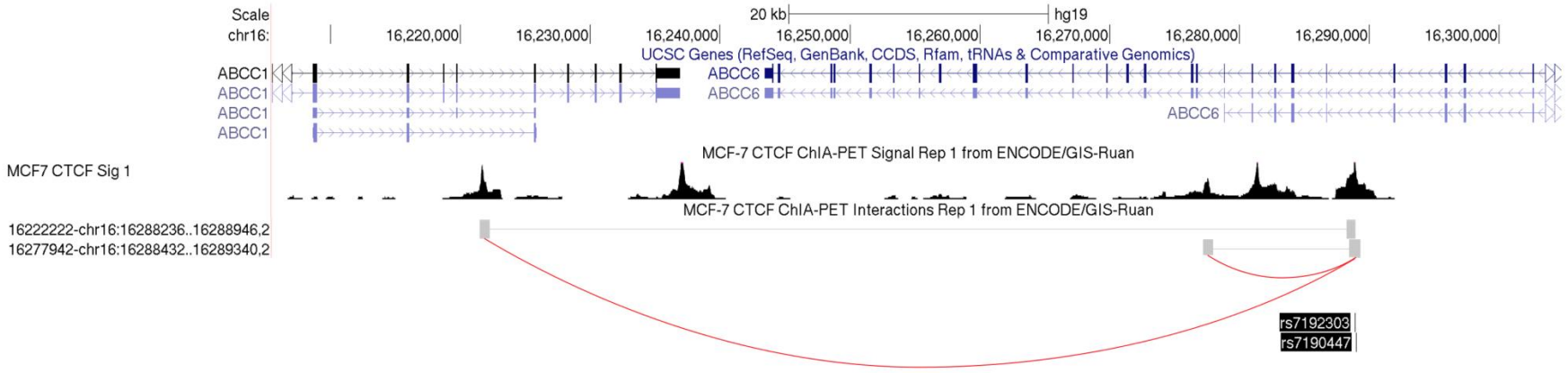


Figure R4. CTCF-mediated chromatin interactions for rs7190447 and rs7192303 (*ABCC6*) determined by chromatin interaction paired-end tag (ChiAPET) data from ENCODE. UCSC Genome Browser image of the genomic region containing rs7190447 (our replicated variant) and rs7192303 (the variant in perfect LD with it) showing ChIA-PET interactions and enrichment for CTCF in MCF-7 breast cancer cells. CTCF-mediated chromatin interactions are represented by two blocks, one at each end, connected by a horizontal line. The density graph shows the CTCF signal enrichment based on aligned read density. Not all MCF-7-CTCF ChIA-PET interactions are shown in full for the genomic region and the chromatin interactions have been adapted (red lines connecting blocks) to highlight relevant interactions only.


A
Scale
chr7:
20 kb
hg19
87,140,000
87,150,000
87,160,000
87,170,000
87,180,000
87,190,000
87,200,000
87,210,000
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
ABCB1
ABCB1
DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3)
DNase Clusters
Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs
Txn Factor ChIP
Simple Nucleotide Polymorphisms (dbSNP 142) Found in >= 1% of Samples
Common SNPs(142)
B
Scale
chr7:
200 bases
hg19
87,170,950
87,171,000
87,171,050
87,171,100
87,171,150
87,171,200
87,171,250
87,171,300
87,171,350
87,171,400
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
ABCB1
ABCB1
DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3)
DNaseI clusters
2
3
Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs
Txn factor binding
RUNX3
G
GM12878 Chromatin State Segmentation by HMM from ENCODE/Broad
13_Heterochrom/lo
11_Weak_Txn
7_Weak_Enhancer
H1-hESC Chromatin State Segmentation by HMM from ENCODE/Broad
13_Heterochrom/lo
K562 Chromatin State Segmentation by HMM from ENCODE/Broad
13_Heterochrom/lo
HepG2 Chromatin State Segmentation by HMM from ENCODE/Broad
Chromatin states
11_Weak_Txn
HUVEC Chromatin State Segmentation by HMM from ENCODE/Broad
13_Heterochrom/lo
HMEC Chromatin State Segmentation by HMM from ENCODE/Broad
13_Heterochrom/lo
HSMM Chromatin State Segmentation by HMM from ENCODE/Broad
13_Heterochrom/lo
NHEK Chromatin State Segmentation by HMM from ENCODE/Broad
13_Heterochrom/lo
Simple Nucleotide Polymorphisms (dbSNP 142) Found in >= 1% of Samples
rs28381937
rs28381936
rs4148737
rs4148736

Figure R5. ENCODE functional evidence displayed in the UCSC Genome Browser for rs4148737 and nearby variants. (**A**) Genomic location of the *ABCB1* gene. Multiple DNase-seq and transcription factor ChIP-seq clusters can be observed. (**B**) Genomic location of rs4148737 (highlighted). Two DNase hypersensitivity regions are observed in five ENCODE cell lines around rs4148737, one of them overlapping. rs4148737 resides in a RUNX3 ChIP-seq cluster in lymphoblastoid cells (identified by G letter). Gray boxes indicate the extent of the hypersensitive region for DNaseI hypersensitivity clusters or transcription factor occupancy, with the darkness of the box proportional to the maximum signal strength observed in any cells contributing to the cluster. The number to the left of a DNaseI box shows how many cells are hypersensitive in the region. Within a ChIP-seq cluster, green highlighting indicates the highest scoring site of a Factorbook-identified canonical motif for the corresponding factor. Chromatin states characterized by combinations of histone marks are also shown in different human cell lines. Each chromatin state is associated with a different segment color. Yellow, weak enhancer; light green, weak transcribed. GM12878, lymphoblastoid cells; K1-hESC, embryonic stem cells; K562, erythrocytic leukemia cells; HepG2, hepatocellular carcinoma cells; HUVEC, umbilical vein endothelial cells; HMEC, mammary epithelial cells; HSMM, skeletal muscle muscle myoblast cells; NHEK, normal epidermal keratinocytes cells. Abbreviations: SNP, single nucleotide polymorphism; Txn, transcription.

2. Results, Study II: identification of genetic variants predictive of susceptibility to capecitabine-induced hand-foot syndrome (CiHFS)

The demographic and clinical characteristics of the discovery (N=166) and replication (N=85) cohorts are shown in **Table R5.**

Table R5. Clinical characteristics of the capecitabine-treated cancer patients

	Discovery (N=166)		Replication (N=85)	
Characteristic	**N**	**%***	**N**	**%***
Age at diagnosis (years)				
Median	63		64	
Range	29-88		36-87	
Sex				
Female	142	86	50	41
Male	23	14	35	59
Missing	1			
Primary diagnosis (tumor type)				
Breast cancer	119	72	27	32
Colorectal cancer	46	28	58	68
Capecitabine treatment regimen				
Standard	146	88	72	85
Continuous	20	12	13	15
CiHFS grade				
Grade 0	78	47	61	72
Grade 3	88	53	24	28
Capecitabine cumulative dose (mg/m^2) (median, range)				
Grade 0	260,166.5 (33,333-1,775,000)		30,1000 (35,000-1,766,800)	
Grade 3[a]	116,673.5 (15,000-455,000)		161,000 (33,600-542,500)	

*Percentages are computed based on the total number of non-missing values. a. Median (and range) capecitabine cumulative dose to the development of grade 3 CiHFS.

Both cohorts included adult breast and colorectal cancer patients treated with capecitabine but the discovery cohort enrolled more breast cancer patients (72% *v* 32%). Gender showed more female patients in the discovery cohort (86%), but more males in the replication cohort (59%). In total, 45% of patients developed grade 3 CiHFS with a median capecitabine cumulative dose at the development of severe toxicity of 134,400 mg/m^2. Over 85% of patients in both cohorts received a standard capecitabine treatment.

Three patients in the discovery cohort failed genotyping (call rate<0.95) and 3 patients were excluded as ethnic outliers based on inspection of plots of the two first principal components, leaving 160 patients (85 cases and 75 controls) for further analysis (**Figure R6**).

Q-Q plot showed no obvious population stratification (**Figure R7**).

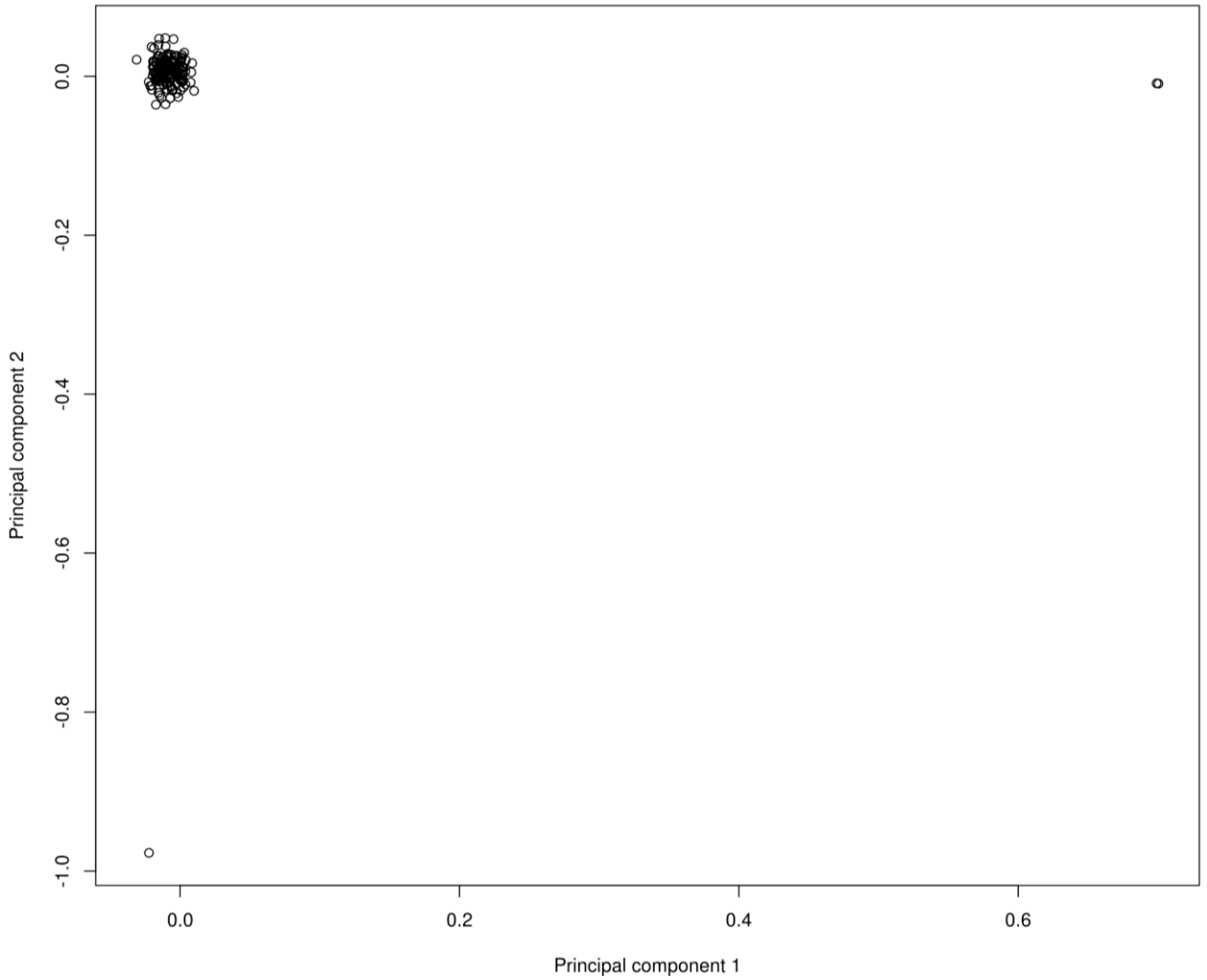


Figure R6. Principal component analysis (PCA) of genetic data for the 163 capecitabine-treated patients from the discovery cohort and variants which passed quality control (520,052). We plotted principal component 1 and principal component 2.

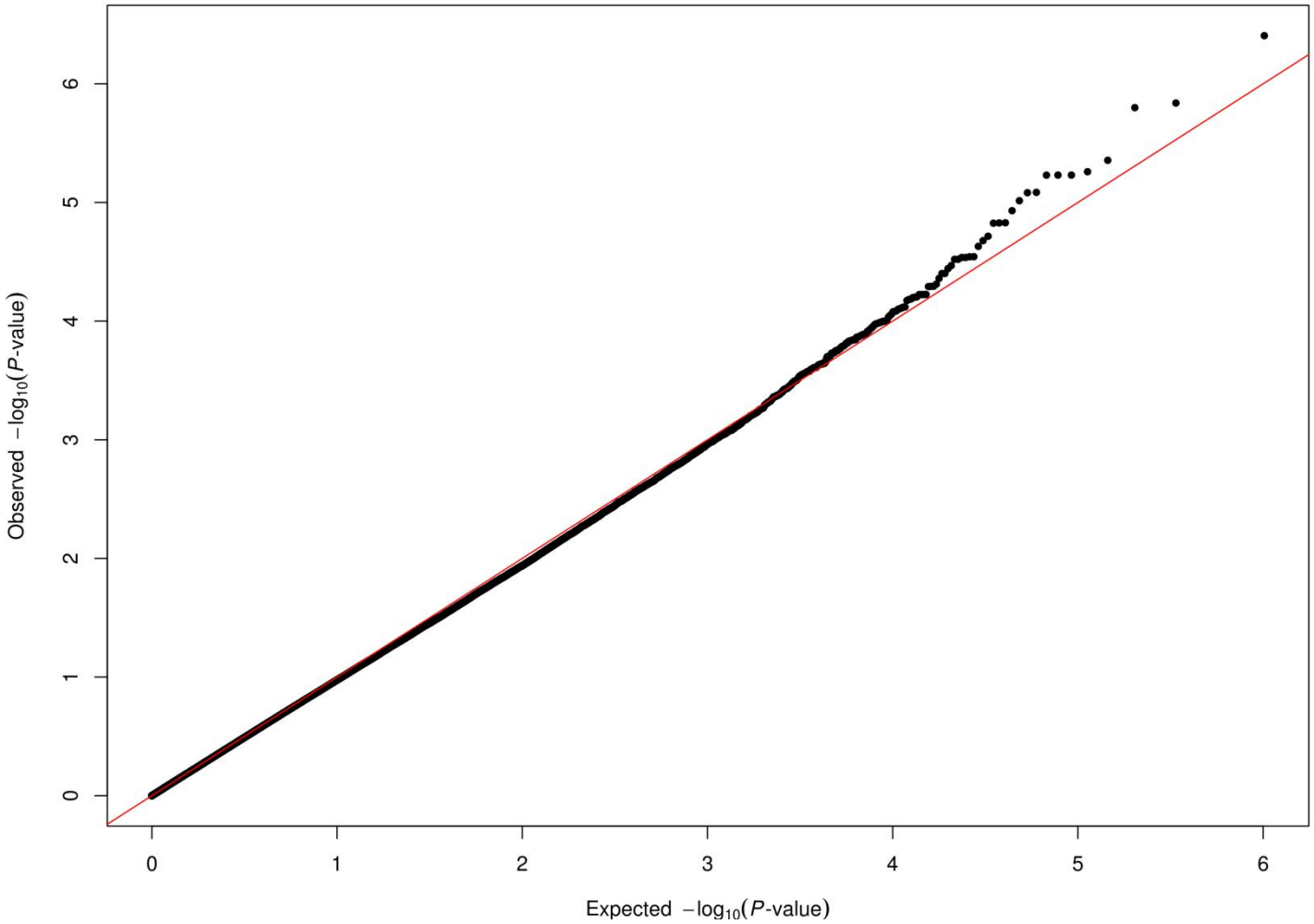


Figure R7. Quantile-quantile (Q-Q) plot. Q-Q plot showing the distributions of observed $-\log_{10}$(*P*-values) from Cox regression analysis adjusted for sex, tumor type and treatment regimen for 520,052 variants which passed QC in the discovery cohort plotted against expected $-\log_{10}$(*P*-values) (λ=0.981). Smaller *P*-values than would be expected by chance are observed at the tail of the plot. λ= 0.981 indicates no obvious population stratification.

2.1. GWAS and fine-mapping

Of variants on the Human610 array, 3,171 failed genotyping (call rate <0.95), 65,496 were excluded by MAF (MAF<0.05), as well as 210 variants whose genotype frequencies departed from Hardy-Weinberg equilibrium ($P<10^{-6}$), leaving a total of 520,052 variants for further analysis.

We carried out single-variant associations using Cox regression analyses, modeling the cumulative dose of capecitabine required for the development of grade 3 CiHFS, and adjusting for gender, tumor type and treatment regimen; and found 10 independent genetic variants that reached a $P<10^{-5}$ (**Figure R8** and **Table R6**). These 10 SNVs were genotyped in the replication cohort (N=85 patients) and associations with grade 3 CiHFS were assessed. The intergenic variant rs6093063 (G>T) was consistently associated with risk of grade 3 CiHFS (**Table R6**), with an estimated HR of 2.27 (95%CI=1.20–4.30, P=0.011). The combined analysis of the 245 subjects from both series gave an HR of 2.16 (95%CI=1.63–2.88, $P=1.32\times10^{-7}$). The Kaplan-Meier analyses showing the association of the intergenic variant rs6093063 (G>T) in both cohorts, modeling the cumulative dose of capecitabine until the development of CiHFS grade 3 is shown in **Figure R9.**

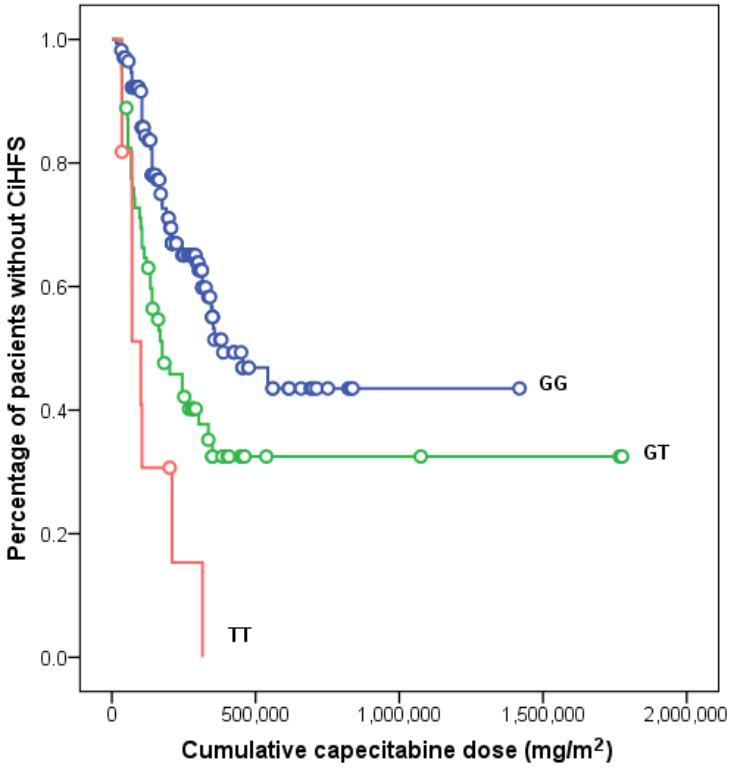


Figure R9. Kaplan–Meier analysis of cumulative dose of capecitabine until the development of severe CiHFS grade 3 according to rs6093063 (G>T) genotypes in the 245 patients from the combined cohort (χ^2=27.06, $P_{log\text{-}rank}=1.3\text{x}10^{-6}$, N_{GG}=170, N_{GT}=63, N_{TT}=11).

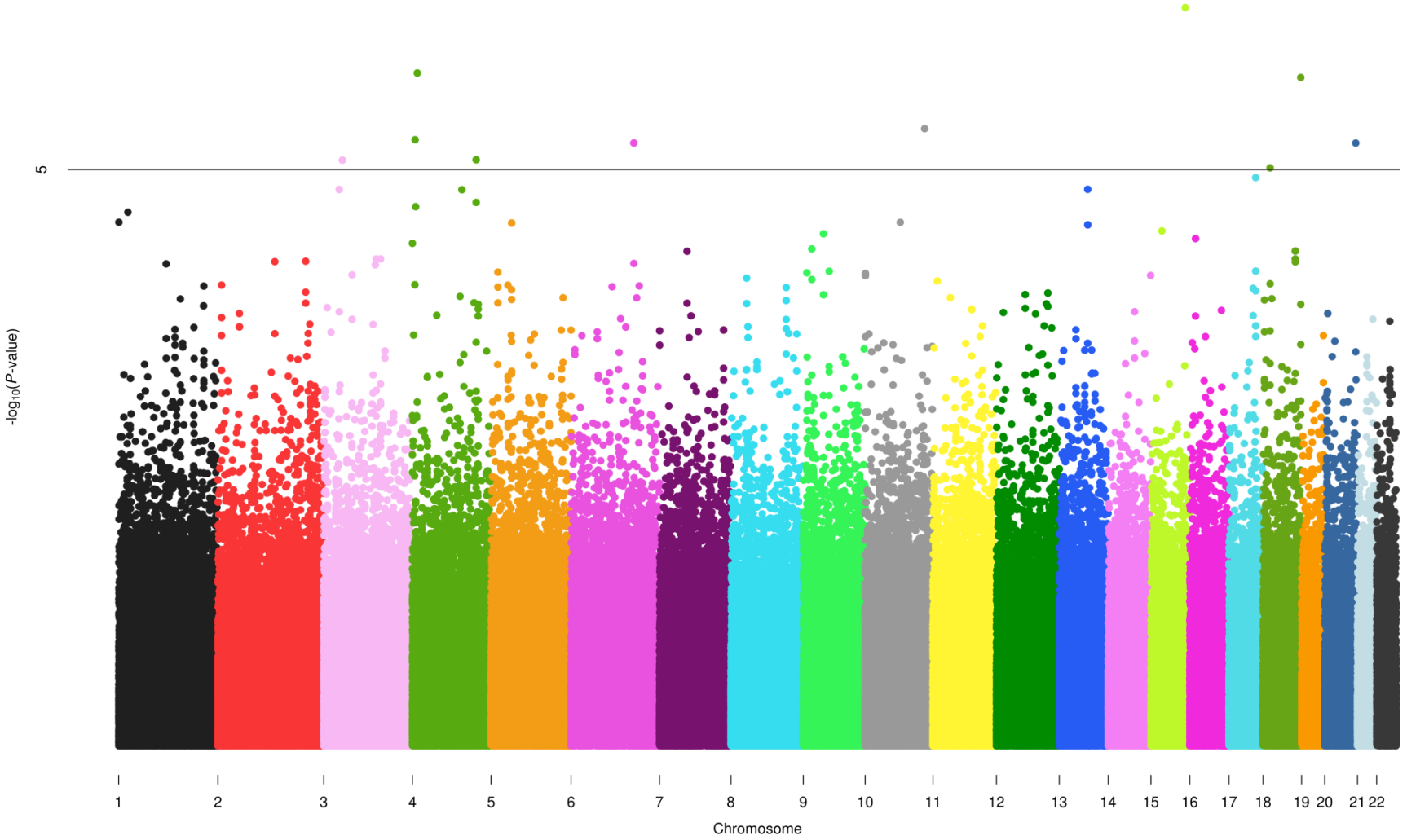


Figure R8. Manhattan plot of the GWAS in the discovery cohort showing association between the genotypes of 520,052 variants and risk of CiHFS. The $-\log_{10}$(*P*-values) from Cox regression analysis adjusted for sex, tumor type and treatment regimen was plotted against its physical chromosomal position. Variants above the black line represent those with a $P < 10^{-5}$.

Table R6. Association analysis results for ten variants with $P<10^{-5}$ in the discovery analysis

Variant	Chr.	Position *	Cohort	MAF	*P*	HR	95%CI
rs7428377	3	30299265	Discovery	0.10	$8.27x10^{-6}$	2.82	1.79-4.46
			Replication	0.05	0.95	0.95	0.21-4.24
			Combined	0.09	$2.99x10^{-6}$	2.47	1.61-3.78
rs10516179	4	6086752	Discovery	0.09	$5.51x10^{-6}$	2.81	1.80-4.40
			Replication	0.14	0.78	0.87	0.34-2.24
			Combined	0.11	$1.23x10^{-3}$	1.97	1.30-2.97
rs12511120	4	8656976	Discovery	0.07	$1.45x10^{-6}$	3.51	2.10-5.87
			Replication	0.09	0.29	1.59	0.66-3.81
			Combined	0.07	$4.33x10^{-5}$	2.3	1.54-3.44
rs1852869	4	160319418	Discovery	0.44	$8.22x10^{-6}$	2.08	1.51- 2.88
			Replication	0.47	0.37	1.38	0.68-2.80
			Combined	0.45	$1.03x10^{-5}$	1.9	1.42- 2.52
rs1121941[a]	6	123241136	Discovery	0.21	$5.88x10^{-6}$	2.46	1.66-3.63
			Replication	0.25	0.44	1.31	0.65-2.63
			Combined	0.22	$2.82x10^{-4}$	1.8	1.31-2.47
rs2420903	10	122802506	Discovery	0.40	$4.41x10^{-6}$	0.47	0.34-0.65
			Replication	0.44	0.76	0.88	0.39-1.98
			Combined	0.41	$9.20x10^{-6}$	0.52	0.39-0.69
rs2619170	15	95213582	Discovery	0.07	$3.93x10^{-7}$	4.12	2.38-7.12
			Replication	0.11	0.38	1.56	0.57-4.24
			Combined	0.08	$1.35x10^{-5}$	2.8	1.76- 4.46
rs7236593	18	11272254	Discovery	0.11	$9.67x10^{-6}$	2.57	1.69-3.92
			Replication	0.12	0.04	0.27	0.07-0.99
			Combined	0.11	$2.83x10^{-3}$	1.82	1.22-2.71
rs13381276	18	76252393	Discovery	0.10	$1.59x10^{-6}$	3.27	2.01-5.31
			Replication	0.20	0.66	1.16	0.58-2.30
			Combined	0.13	$4.47x10^{-4}$	1.86	1.31-2.62
rs6093063	20	59723157	Discovery	0.17	$5.88x10^{-6}$	2.17	1.55-3.04
			Replication	0.18	0.01	2.27	1.20-4.30
			Combined	0.17	$1.32x10^{-7}$	2.17	1.63-2.88

HRs are per copy of the specified minor allele. The replicated variant rs6093063 is shown in bold. SNVs were genotyped in the replication cohort by using KasPar assays. * Chromosome positions are based on Genome Reference Consortium Human Build 37 (GRCh37/hg19). a. rs1121941 is in total LD (r^2=1) with rs4962257. Abbreviations: Chr., chromosome.

Variant rs6093063 is located in a LD block of 38,501 bp spanning the chromosome 20 positions 59,707,264-59,745,765 (GRCh37/hg19) at the 20q13.33 locus. In order to fine-map the rs6093063 association with CiHFS we subsequently genotyped 34 SNVs and imputed 80 additional variants in the 245 subjects from both cohorts and tested their associations with CiHFS. Of the 10 variants with $P<10^{-5}$ in the combined analysis of these 114 (**Table R7**), rs6129058 showed the strongest evidence of association with grade 3 CiHFS (HR=2.40; 95%CI =1.78–3.25; $P=1.2x10^{-8}$) (**Table R7, Table R8,** and **Figure R10**).

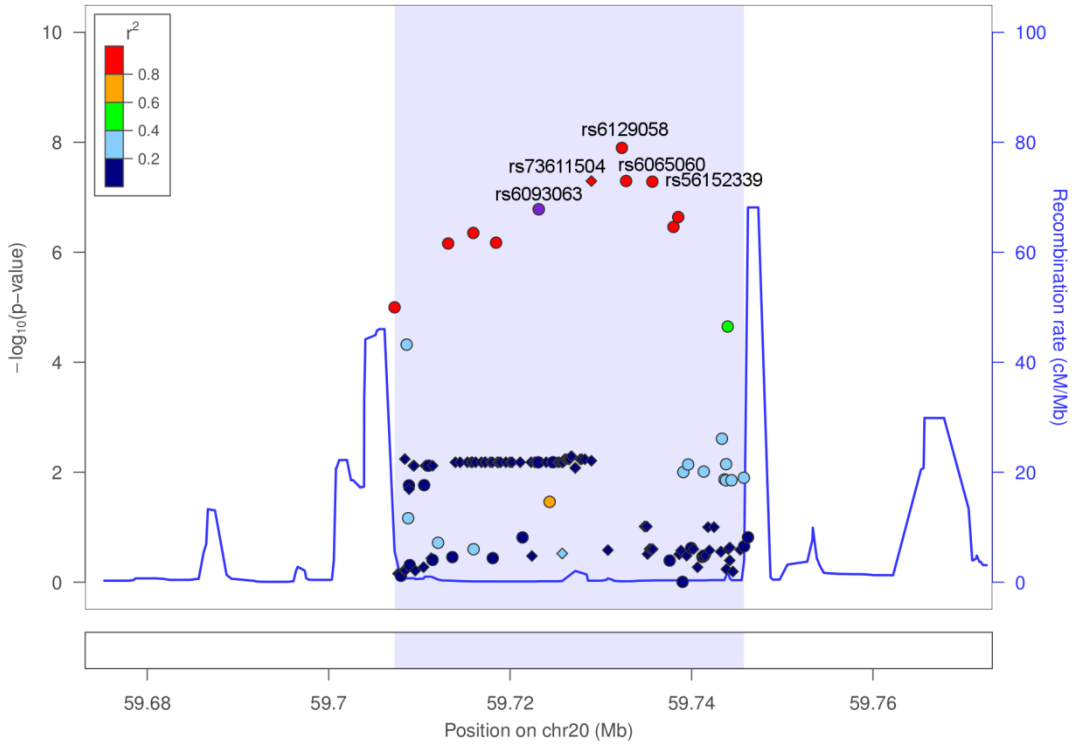


Figure R10. Association and recombination plot of the LD block containing the replicated variant rs6093063 (shaded region) and ±50kb boundaries. Plot shows the genomic region associated with CiHFS and the $-\log_{10}$ association *P*-values of genotyped and imputed variants (MAF>5%). Circles and diamonds indicate genotyped and imputed variants, respectively. Recombination rates are also shown. Variant colour indicates the strength of LD (r^2) with rs6093063. Recombination rates are based on 1000 Genomes Project and genomic coordinates are based on Genome Reference Consortium Human Build 37 (GRCh37/hg19).

We next conducted haplotype analyses based on the four most strongly CiHFS-associated variants ($P<5x10^{-7}$; **Table R8,** and **Figure R10**), which revealed one rare (h1) and one more common (h2) haplotypes (**Table R9**). The rare haplotype (h1) carried the risk alleles at rs6129058, rs56152339 and rs6065060, but not at rs6093063. The HR estimate for this haplotype relative to the reference haplotype (no risk alleles) was lower compared to that for h2 (HR h1=1.82 *v* HR h2=2.48), which carried the risk alleles in all four variants, but the confidence intervals were entirely overlapping. Haplotype analyses suggest that the four novel risk variants associated with susceptibility to CiHFS could be causal, either individually or in combination.

2.2. Risk haplotype is associated with reduced CDH4 mRNA expression

To explore how the identified variants confer susceptibility to CiHFS, we first examined the genomic landscape of the risk variant-containing 20q13.33 locus. The closest gene, located ~90 kb away, is *CDH4* and encodes R-cadherin. Cadherin family members are master regulators of the cohesiveness of epithelial layers, and genetic defects in cadherin genes have been linked to inherited skin disorders[183–187]. Although R-cadherin has not been previously implicated in skin physiology, immunohistochemical approaches revealed that R-cadherin mainly localized in the

suprabasal granular layer of the stratified human epidermis (**Figure R11A** and **Figure R12B**). This distribution differed from that documented for other cadherin family members (**Figure R11B** and **Figure R12**). This result argued for a potential involvement of *CDH4* gene in CiHFS.

Table R7. Strongest associated variants ($P<10^{-5}$) associated with CiHFS appearance after fine-mapping

Variant	Chr.	Position *	Type	*P*	HR	95%CI
rs6071509	20	59713172	Genotyped	$1.01x10^{-6}$	2.12	1.57 - 2.87
rs68114568	20	59715943	Genotyped	$6.46 x10^{-7}$	2.11	1.575 - 2.84
rs1074095	20	59718453	Genotyped	$6.46 x10^{-7}$	2.11	1.57 - 2.84
rs6093063	20	59723157	Genotyped	$1.32 x10^{-7}$	2.17	1.63 - 2.88
rs73611504	20	59728958	Imputed	$5.07 x10^{-8}$	2.32	1.72 - 3.15
rs6129058	20	59732322	Genotyped	$1.17 x10^{-8}$	2.40	1.78- 3.25
rs56152339	20	59732787	Genotyped	$7.38 x10^{-8}$	2.30	1.70 - 3.12
rs6065060	20	59735681	Genotyped	$7.59 x10^{-8}$	2.30	1.70- 3.12
rs6101248	20	59738006	Genotyped	$1.39 x10^{-6}$	2.14	1.57- 2.92
rs6071520	20	59738531	Genotyped	$1.13 x10^{-6}$	2.23	1.62 - 3.08

Additive model of inheritance was considered. Type indicates whether SNVs were genotyped or imputed. HRs are per copy of the specified minor allele. * Chromosome positions are based on Genome Reference Consortium Human Build 37 (GRCh37/hg19). Abbreviations: Chr., chromosome.

Table R8. Association results for the two stages and the combined analysis at the 20q13.33 locus for the 3 most strongly associated variants and the original GWAS hit**

Variant	Position *	Cohort	MAF	*P*	HR	95%CI
rs6093063**	20:59723157	Discovery	0.17	$5.9 x10^{-6}$	2.17	1.55-3.04
		Replication	0.18	0.011	2.27	1.20-4.30
		Combined	0.17	$1.3 x10^{-7}$	2.17	1.63-2.88
rs6129058	20:59732322	Discovery	0.18	$1.3 x10^{-6}$	2.46	1.71-3.54
		Replication	0.21	$2.9 x10^{-3}$	2.70	1.40-5.19
		Combined	0.19	$1.2 x10^{-8}$	2.40	1.78-3.25
rs56152339	20:59732787	Discovery	0.18	$4.8 x10^{-6}$	2.35	1.63-3.39
		Replication	0.19	$6.8 x10^{-3}$	2.45	1.28-4.69
		Combined	0.19	$7.4 x10^{-8}$	2.30	1.70-3.12
rs6065060	20:59735681	Discovery	0.18	$4.8 x10^{-6}$	2.35	1.63-3.40
		Replication	0.20	$6.9 x10^{-3}$	2.49	1.28-4.69
		Combined	0.19	$7.6 x10^{-8}$	2.30	1.70-3.12

Additive model of inheritance was considered. HRs are per copy of the minor allele. * Chromosome positions are based on Genome Reference Consortium Human Build 37 (GRCh37/hg19). ** rs6093063.

Table R9. Haplotype analyses including the four most strongly associated variants

Haplotypes	rs6093063 (G>T)	rs6129058 (G>T)	rs56152339 (T>A)	rs6065060 (T>A)	Haplotype frequency	*P*	HR	95%CI
Reference [GGTT]	0	0	0	0	0.80	-	-	
1 [GTAA]	0	1	1	1	0.02	0.24	1.82	0.67-4.92
2 [TTAA]	1	1	1	1	0.17	$1.43x10^{-8}$	2.48	1.81-3.39
All Others			Rare		0.01	0.48	1.51	0.48-4.76

Each haplotype was compared to the ancestral haplotype carrying the common alleles of all 4 variants (reference). Variants rs56152339 and rs6065060 were almost perfectly correlated with each other

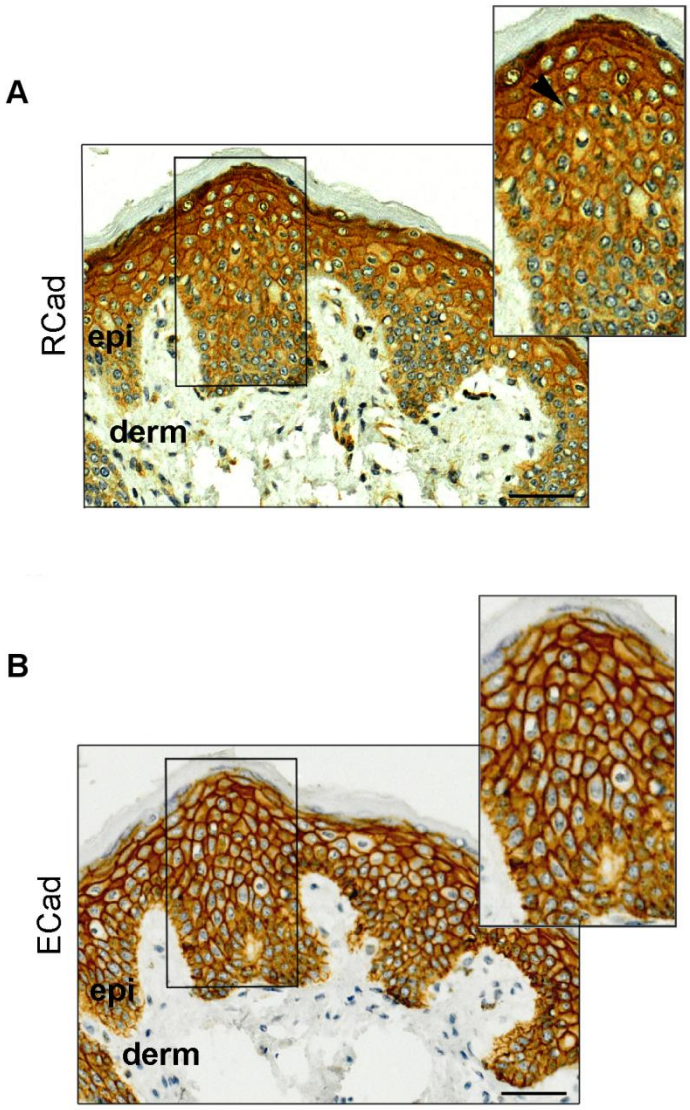


Figure R11. R-cadherin distribution in the stratified human epidermis. Immunohistochemical analyses of (**A**) R-cadherin (RCad) and (**B**) E-cadherin (ECad) in human control skin. A magnification of the selected areas is shown. Arrowhead indicates localization of R-cadherin at cell-cell contacts in suprabasal layers of the epidermis. Scale bar, 50 µm. Abbreviations: epi, epidermis; derm, dermis.

eQTL analyses in normal human tissues samples revealed a significant association between the presence of the risk haplotype (h2) and reduced *CDH4* mRNA expression compared to the reference haplotype (*P*=0.017; **Figure R13**).

2.3. The risk allele containing locus interacts with the CDH4 promoter

ChIP-seq data available in ENCODE for normal human epidermal keratinocytes (NHEK) showed no significant transcription factor binding or defined enhancer overlapping with our risk variants. However, binding sites of CTCF and cohesin were present close to the locus (**Figure R14A, top**). As these factors are involved in chromatin

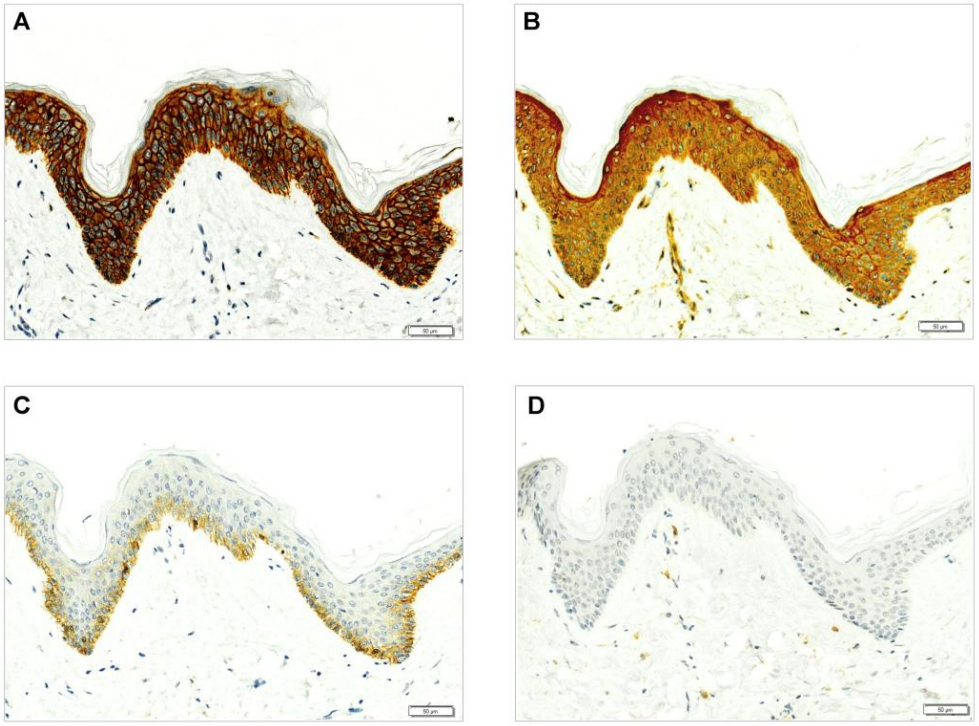


Figure R12. Immunohistochemical analyses of four cadherins. A) E-cadherin, **B**) R-cadherin, **C**) P-cadherin and **D**) N-cadherin in normal skin samples from breast cancer patients prior capecitabine treatment. R-cadherin was found to have a different localization in skin tissue (highly expressed in the suprabasal granular layers of the epidermis) compared to E-cadherin (strongly expressed throughout the entire epidermis) and P-cadherin (only expressed in the basal layers of the epidermis). N-cadherin is not expressed in skin and was used as a negative control. Scale bar, 50 µm.

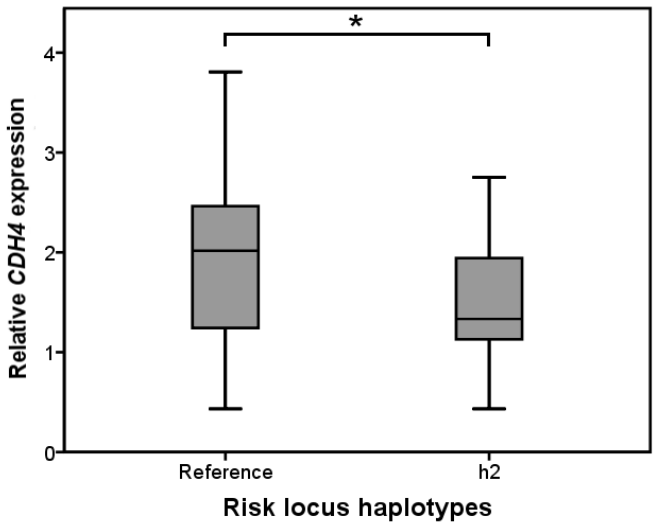


Figure R13. Association between risk locus haplotypes and *CDH4* mRNA expression levels. Boxplot of the *CDH4* mRNA levels normalized to β-actin in 44 normal human livers. Natural log transformed expression levels are represented. Differences in mRNA levels were assessed using the Student's t test ($N_{reference}$=67, N_{h2}=18, *P*=0.017). *indicates *P*<0.05. Haplotypes h1 (N=1) and "others" (N=2) were not included in the analysis.

folding and DNA looping[188–190], we hypothesized that regulation of *CHD4* expression may require specific chromatin contacts involving the risk locus. To test this, we employed circularized chromosome conformation capture, followed by high-throughput sequencing analyses (4C-seq) in keratinocyte cell lines homozygous for the reference and h2 haplotypes, setting the viewpoint at the *CDH4* promoter. By means of 4C-seq we could detect chromatin contacts between the risk locus and the *CDH4* promoter in keratinocytes homozygous for the risk alleles (**Figure R14A**, **bottom**). Interestingly, the interaction was much less apparent in another keratinocyte cell line homozygous for the reference haplotype (**Figure R14B**). Similar results were obtained using a second viewpoint at the *CDH4* promoter (**Figure R15**). These results suggest a mechanism by which the presence of the risk alleles can affect *CDH4* gene expression. Taken together the results from eQTL and 4C-seq analyses, we concluded that *CDH4* is a strong candidate to be involved in CiHFS.

2.4. CDH4-deficiency leads to decreased levels of involucrin

To explore the contributions of *CDH4* gene to the proper differentiation of keratinocytes, we next turned to keratinocytes cultured in vitro, both under proliferative conditions and upon calcium-induced differentiation. Proliferative keratinocytes are normally found in the basal layers of the epidermis and move to suprabasal layers upon differentiation. In culture, keratinocytes proliferate and grow in a monolayer in low calcium medium but the addition of calcium promotes cell-cell contact formation and differentiation into postmitotic keratinocytes[158]. So, keratinocytes were switched to a calcium containing media to induce differentiation and expression of R-cadherin was evaluated at different time points of the calcium switch and compared to low calcium conditions. Immunoblot analyses revealed that R-cadherin was highly expressed upon differentiation (**Figure R16**), consistent with its primary localization at the suprabasal layer of the epidermis (**Figure R11A** and **Figure R12B**). This increase in R-cadherin expression was more noticeable after 12 hours of calcium addition.

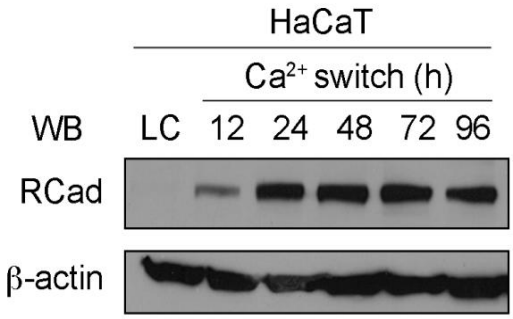


Figure R16. R-cadherin protein levels in human keratinocytes. Western blot (WB) analysis of R-cadherin (RCad) levels in HaCaT keratinocytes both in low calcium conditions (LC) and at different time points of a calcium switch time course.

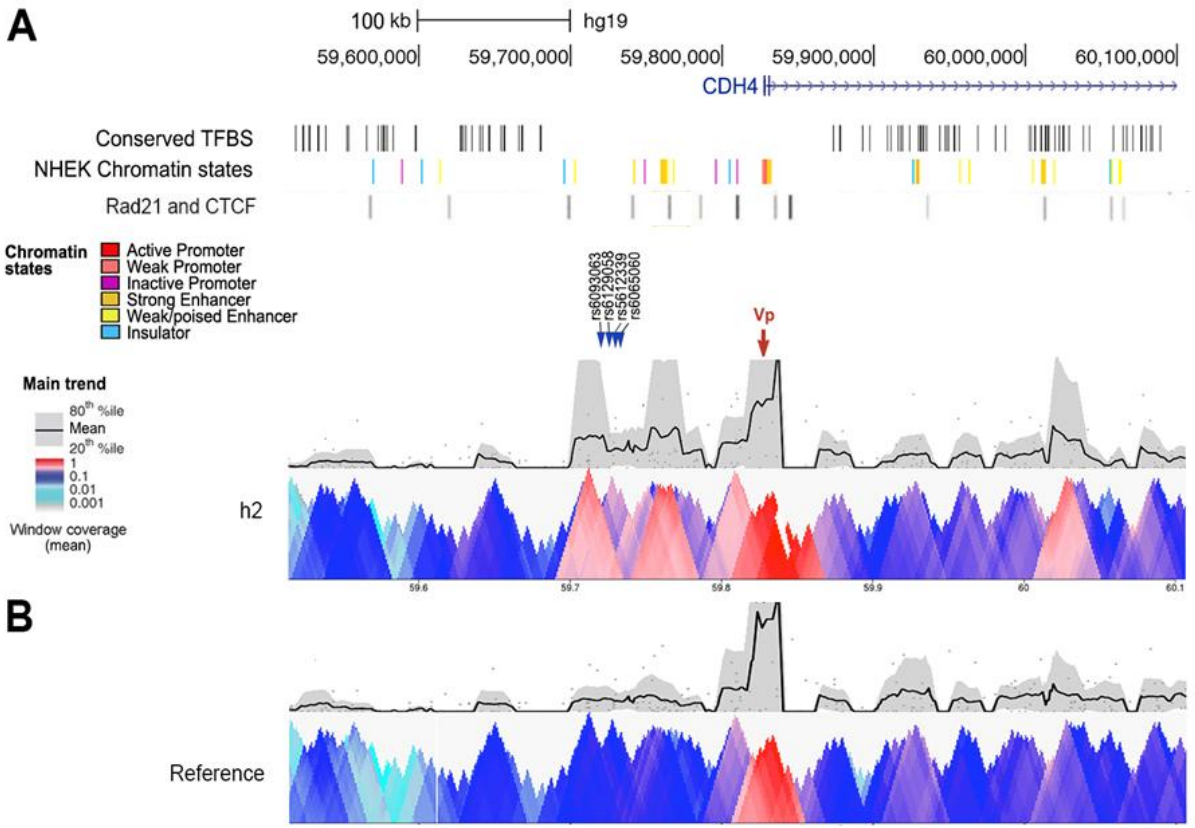


Figure R14. Chromatin interactions at the 20q13.33 between the *CDH4* promoter and the CiHFS susceptibility locus. (**A**) Detail of the 20q13.33 CiHFS susceptibility locus located ~ 90kb upstream of the *CDH4* gene promoter showing, from the top down, chromosomal coordinates, a track for HMR conserved transcription factor binding sites (TFBS), a track with chromatin states defined in NHEK cells from ENCODE/Broad (colour code appears below), the binding sites for CTCF and cohesin Rad21 from ChIP-seq experiments from ENCODE, the positions of the four most strongly associated variants (blue arrowheads) and the contact profile from a viewpoint located in the *CDH4* gene promoter (Vp, red) in cells homozygous for the risk haplotype (h2). This profile was generated using a 20 kb window size in the main trend subpanel. High contact probability (red) is observed between the *CDH4* promoter and the region containing CiHFS associated variants. (**B**) Using the same viewpoint in a different keratinocyte cell line homozygous for the reference haplotype, this interaction is clearly decreased. Similar results were obtained with a second viewpoint around the *CDH4* promoter (**Figure R15**).

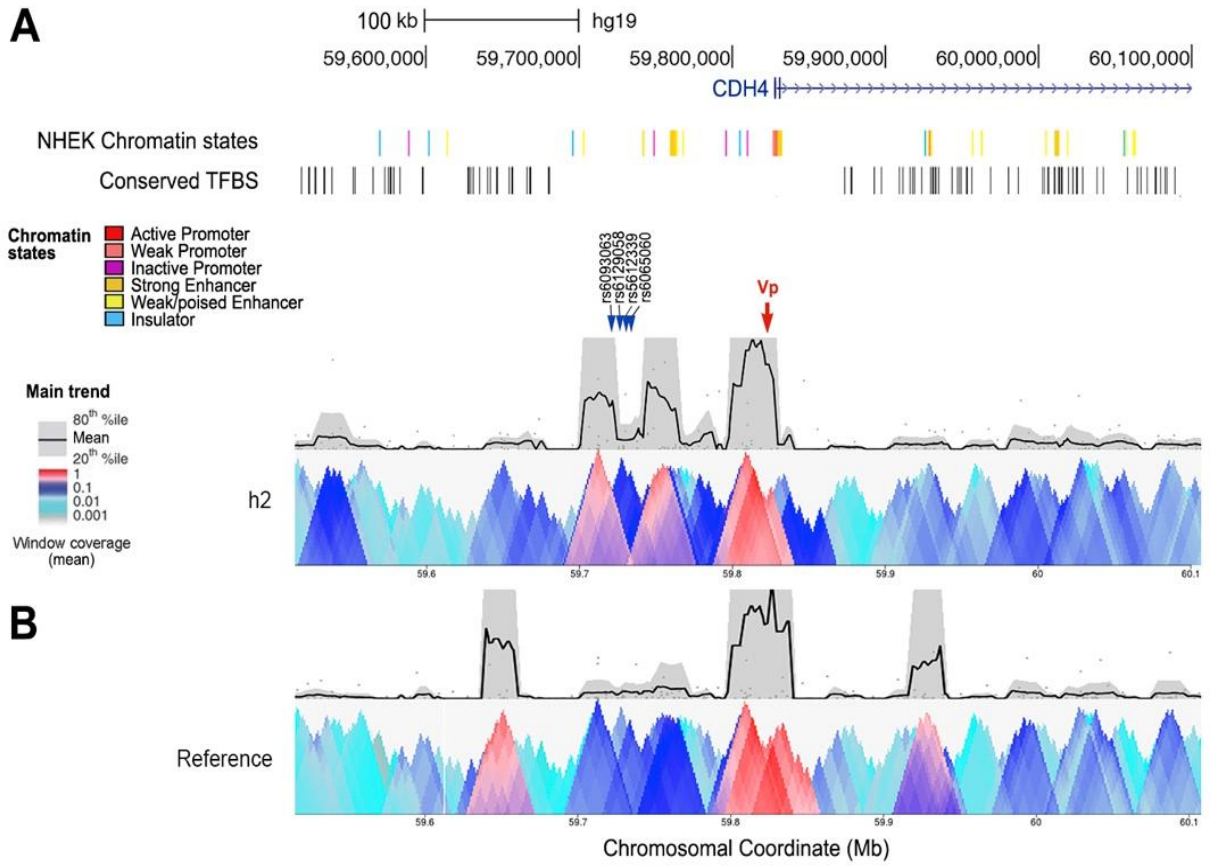


Figure R15. Chromatin interactions at the 20q13.33 between the *CDH4* promoter and the CiHFS susceptibility locus were validated with an independent viewpoint. Representation of the 20q13.33 CiHFS susceptibility locus showing, from the top down, chromosomal coordinates, a track for HMR conserved transcription factor binding sites (TFBS), a track with chromatin states defined in NHEK cells from ENCODE/Broad (colour code appears below), the positions of the four most strongly associated variants (blue arrowheads) and the contact profile from a second viewpoint located close to the *CDH4* gene promoter (Vp, red arrow) in cells homozygous for the risk haplotype (**A**) or for the reference haplotype (**B**), as in **Figure R14**.

We next assessed the consequences of decreasing R-cadherin in keratinocyte differentiation using a specific short hairpin RNA (shRNA) expressed in a retroviral vector. The knock-down was efficient, as observed by qRT-PCR (**Figure R17A**) and western blot (**Figure R17B**). Importantly, E-cadherin mRNA and protein levels did not change after knocking-down *CDH4* (**Figure R17A** and **R17B**). Of note, keratins 5 and 14 are the main structural protein products in proliferating basal keratinocytes; while upon differentiation epidermal keratinocytes express differentiation markers of suprabasal layers such as keratin 1 , keratin 10 filaggrin, loricrin and involucrin[191]. So next, we evaluated the levels of the above mentioned markers of proliferation and terminal epidermal keratinocytes differentiation by qRT-PCR and by immunoblot analyses.

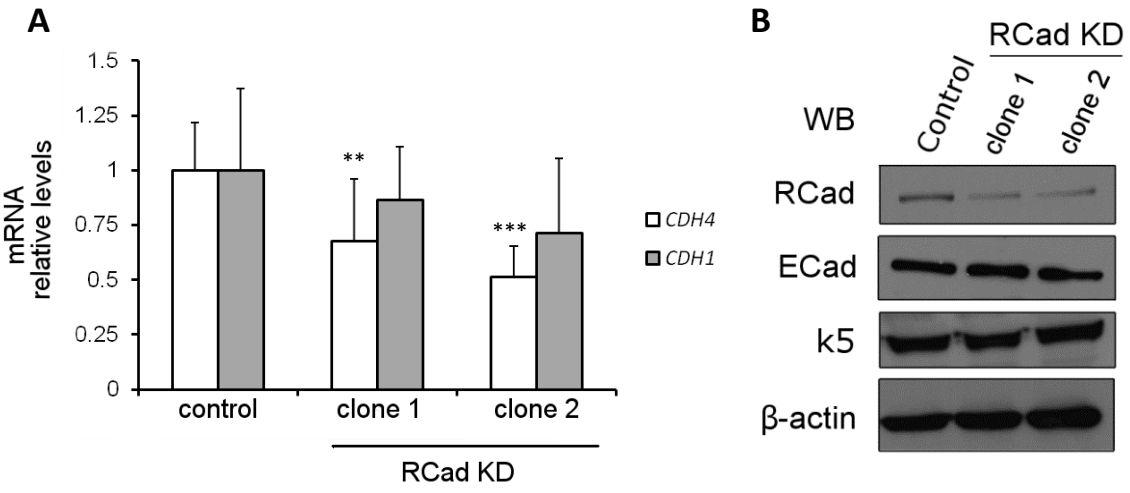


Figure R17. Consequences of *CDH4*-deficiency for *CDH1* (E-cadherin) and keratin 5 (k5) expression. (**A**) qRT-PCR analyses of *CDH4* (R-cadherin) and *CDH1* (E-cadherin) mRNA levels in scramble control cells and two different *CDH4* KD clones after 48 h of calcium treatment. Data was normalized to control values and represented as mean ± s.e.m and assessed using the Student's t test (N=5 independent experiments). ** $P<0.005$, *** $P<0.0001$. (**C**) Western blot (WB) analyses of R-cadherin (RCad), E-cadherin (ECad) and keratin 5 (K5) in scramble control cells and two *CDH4* KD clones after 48 h of calcium treatment. Abbreviations: s.e.m, standard error of the mean.

We found that *CDH4*-deficiency did not have an effect on the expression of keratin 5 (**Figure R17B**), measured 48 hours after calcium addition. In addition, we observed no changes in the expression of keratin 1, keratin 10, filaggrin or loricrin (**Figure R18A** and **R18C**). However, both mRNA (**Figure R18B**) and protein levels (**Figure R18C**) of involucrin were clearly reduced in the two stable *CDH4* knock-down clones compared to control cells.

2.5. R-cadherin and involucrin levels in skin are inversely correlated with risk of CiHFS

To assess the implications of these findings in the susceptibility of CiHFS, we next evaluated by immunohistochemistry the association of the risk alleles identified in our series with the expression of R-cadherin and involucrin in the epidermis from breast cancer patients sampled

prior capecitabine treatment. The presence of the risk alleles correlated with reduced levels of R-cadherin (**Figure R19A**), in agreement with *CDH4* mRNA expression; and involucrin (**Figure R19B**).

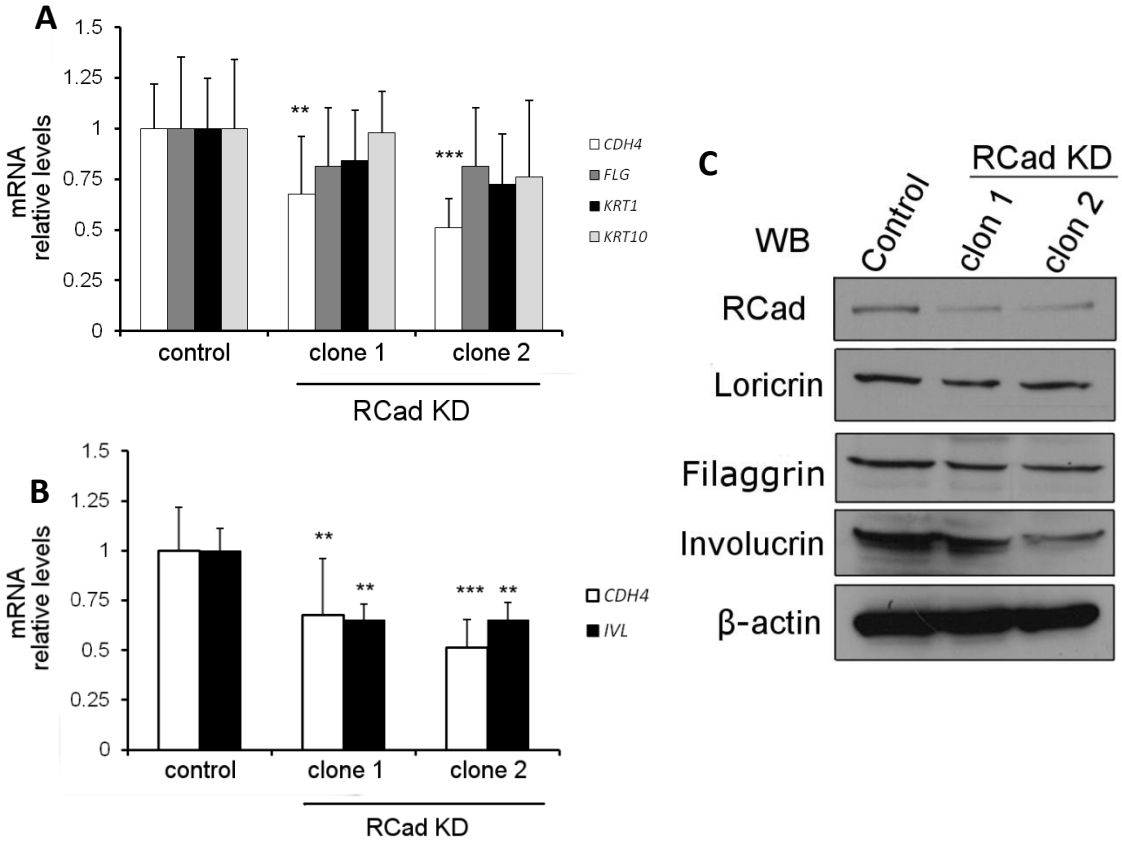


Figure R18. Consequences of *CDH4*-deficiency for the expression of differentiation markers keratin 1, keratin 10, filaggrin, loricrin and involucrin. (**A**) qRT-PCR analyses of *CDH4* (R-cadherin), *FLG* (filaggrin), *KRT1* (keratin 1), *KRT10* (keratin 10) and (**B**) *IVN* (involucrin) mRNA levels in scramble control cells and two different *CDH4* KD clones after 48 h of calcium treatment. Data was normalized to control values and represented as mean ± s.e.m and assessed using the Student's t test (N=5 independent experiments). ** $P<0.005$, *** $P<0.0001$. (**C**) Western blot (WB) analyses of R-cadherin (RCad), loricrin, filaggrin and involucrin in scramble control cells and two different *CDH4* KD clones after 48 h of calcium treatment. Abbreviations: s.e.m, standard error of the mean.

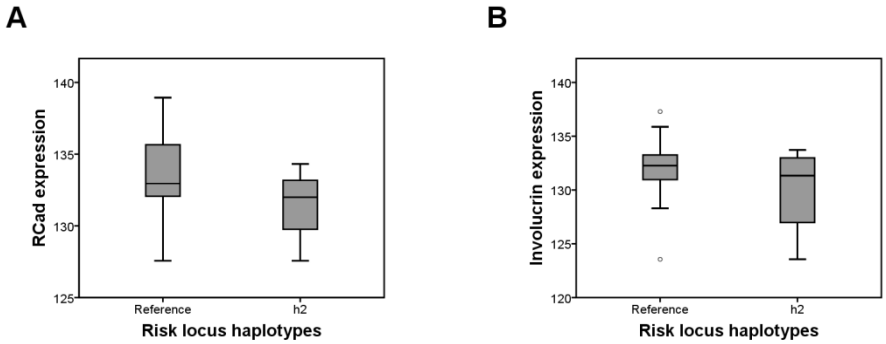


Figure R19. Boxplot of the R-cadherin and involucrin expression in the epidermis from 16 breast cancer patients prior capecitabine treatment. R-cadherin (RCad) (**A**) and involucrin (**B**) expression was quantified as positive immunohistochemically staining intensity in suprabasal layers of 16 skin samples from breast cancer patients prior capecitabine treatment suffering CiHFS grade 0 (N=4) or CiHFS grade 3 (N=12). Skin samples were taken before capecitabine treatment. Differences in R-cadherin and involucrin expression between the reference and risk h2 haplotypes were assessed using the Student's t test ($N_{reference}=23$, $N_{h2}=4$, $P_{R\text{-}cadherin}=0.156$, $P_{involucrin}=0.224$). Haplotypes h1 (N=1) and "others" haplotypes (N=2) were not included in the analysis.

Moreover, we also observed an association of the expression of R-cadherin and involucrin with the risk of suffering this adverse event, wherein before starting the treatment, the skin of the patients exhibiting severe toxicity (grade 3), expressed lower R-cadherin and involucrin levels compared to that of patients suffering no toxicity (grade 0) (**Figure R20A** and **20B**, respectively).

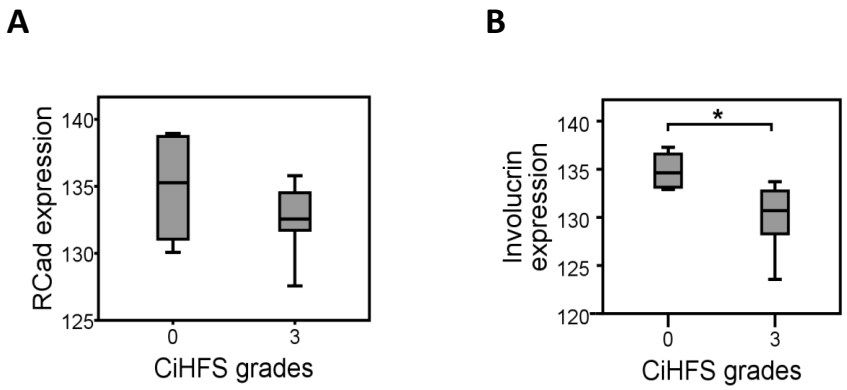


Figure R20. Boxplot of the R-cadherin and involucrin expression in the epidermis from 16 breast cancer patients prior capecitabine treatment. R-cadherin (RCad) (**A**) and involucrin (**B**) expression was quantified as positive immunohistochemically staining intensity in suprabasal layers of 16 skin samples from breast cancer patients prior capecitabine treatment suffering CiHFS grade 0 (N=4) or CiHFS grade 3 (N=12). Skin samples were taken before capecitabine treatment. Differences in R-cadherin and involucrin expression between CiHFS grades 0 and 3 were assessed using the Student's t test ($P_{\text{R-cadherin}}$=0.221; *$P_{\text{involucrin}}$ =0.001).

Main results Study II

We discovered and replicated a cluster of four highly correlated variants associated with susceptibility to CiHFS at 20q13.33 locus (top hit=rs6129058, HR=2.40, 95%CI=1.78–3.20; P=1.17x10^{-8}). Using 4C-sequencing, we identified a chromatin contact between the locus containing the risk alleles and the promoter of *CDH4*, located 90 kb away. The risk haplotype was associated with decreased levels of *CDH4* mRNA and the protein it encodes, R-cadherin, which mainly localizes in the granular layer of the epidermis. In human keratinocytes, *CDH4* downregulation resulted in reduced expression of involucrin, while there were no changes in the expression of others differentiation markers such as, keratin 1, keratin 10, filaggrin or loricrin. Interestingly, immunohistochemical analyses revealed that skin from patients with severe CiHFS exhibited low levels of R-cadherin and involucrin before capecitabine treatment.

3. Results, Study III: identification of genetic variants predictive of susceptibility to chronic anthracycline-induced cardiotoxicity (AIC) in pediatric oncology patients

The demographic and clinical characteristics of the 93 Spanish anthracycline-treated pediatric cancer patients are shown in **Table R10**.

Table R10. Clinical characteristics of the anthracycline-treated pediatric cancer patients

	Controls (N=58)		Cases (N=35)		
Characteristic	**N**	**%***	**N**	**%***	***P***
Age at diagnosis (years)					**0.004**
Median	5.1		10.4		
Range	1.4-16.9		1.2-21.1		
Sex					0.19
Female	25	43	10	29	
Male	33	57	25	71	
Primary diagnosis (tumor type)					
Leukemia	51	88	13	37	**<0.001**
Osteosarcoma	3	5.2	12	34	**<0.001**
Ewing Sarcoma	4	6.9	10	29	**0.007**
Radiotherapy involving the heart[a]	3	6.7	6	19	0.15
Cumulative anthracycline dose (mg/m^2)[b]					**<0.001**
Mean	130		360		
Range	49.2-562		105-780		
≤ 200	43	74	8	23	
> 200	15	26	27	77	
Anthracycline type					
Doxorubicin	50	86	34	97	0.31
Daunorubicin	8	14	1	2.9	0.15
Epirubicin	5	8.6	-	-	0.15
Concomitant therapy					
Cyclophosphamide	55	95	27	77	0.06
Vincristine	54	93	32	91	0.42
Etoposide	14	24	13	37	0.24
Bleomycin	2	3.4	13	37	**<0.001**
Follow-up (years)					**0.06**
Median	8.3		10.5		**0.06**
Range	1-24.1		1-27.5		

Age, cumulative anthracycline dose and follow-up were analyzed by Wilcoxon-Mann-Whitney U test. Sex, tumor type, radiotherapy involving the heart, anthracycline type and concomitant therapy were analyzed by Fisher's exact test. * Percentages are computed based on the total number of non-missing values. a Radiotherapy involving the heart includes mediastinal and mantle radiation and total body irradiation. b Cumulative anthracycline dose was calculated using doxorubicin equivalents. Bold fold indicates statistically significant *P*-values (*P*<0.05).

Controls were significantly younger than cases at diagnosis (median age 5.1 and 10.4 years, respectively, *P*=0.004) and gender distribution in our cohort showed more female patients in controls than in cases (43% *v* 29%); although patients with younger ages and women are significantly particularly vulnerable to AIC[104]. There were fewer cases than controls diagnosed with leukemia (37% v 88%; *P*<0.001) but more with pediatric bone tumors (34% v 5.2% with osteosarcoma; *P*<0.001 and 29% *v* 6.9% with Ewing sarcoma; *P*=0.007). Cumulative anthracycline

dose was higher in cases than in controls (360 mg/m^2 *v* 130 mg/m^2 *P*<0.001), with doxorubicin being the most frequent anthracycline drug administered. Remarkably, patients diagnosed with pediatric bone tumors received higher cumulative anthracycline doses than those diagnosed with leukemia [median cumulative anthracycline dose (mg/m^2)= 439 *v* 132, respectively]. Regarding concomitant therapy, bleomycin was more often administered in cases than in controls (37% *v* 3.4%; *P*<0.001) and there was a trend toward a significant difference in use of cyclophosphamide as concomitant therapy (95% *v* 77%; *P*=0.06).

Three patients (1 case and 2 controls) failed genotyping (call rate<0.95) and 7 patients (4 cases and 3 controls) were excluded as ethnic outliers based on inspection of plots of the two first principal components, leaving 83 patients for further analysis (**Figure R21**).

3.1. Single-variant associations

Of the 247,870 variants on the array, 246,060 passed quality control and 53,136 were polymorphic. Age at diagnosis, cumulative anthracycline dose and bleomycin concomitant administration were included as covariates for logistic regression analyses. After verifying that the association of tumor type with AIC was exclusively explained by the cumulative anthracycline dose, this clinical variable was not included as covariate. The strongest evidence of association was found for the variant rs858345 located in the *ENPP1* (ectonucleotide pyrophosphatase/phosphodiesterase 1) gene at chromosome 6 (P=2.79x10^{-4}, OR=11.3) (**Figure R22**). As could be expected, given the small number of patients relative to the number of genetic variants tested, this variant did not survive a correction for multiple comparisons (P_{FDR}=0.96).

3.2. Gene-based associations

We then carried out gene-based tests by using SKAT-O[55,166] to further investigate the join effects of variants within each gene and considering age at diagnosis, cumulative anthracycline dose and bleomycin concomitant administration as covariates. In total, we tested 4,883 genes of 17,677 covered by at least one variant on the array.

We identified *GPR35* (G protein-coupled receptor 35) as the gene most significantly associated with chronic AIC (P=7.0x10^{-6}) in pediatric patients and this association remained statistically significant after correction for multiple testing (corrected P_{FDR}=0.03). While the SKAT-O does not provide any parameter estimates and in order to evaluate the contribution of each *GPR35* variant, we removed one variant at a time and re-calculated the association for the gene

using SKAT-O. Sensitivity analyses revealed variant rs12468485 (p.Thr253Met; c.758C>T) (MAF=0.04) made the greatest contribution to the observed association (**Figure R23**).

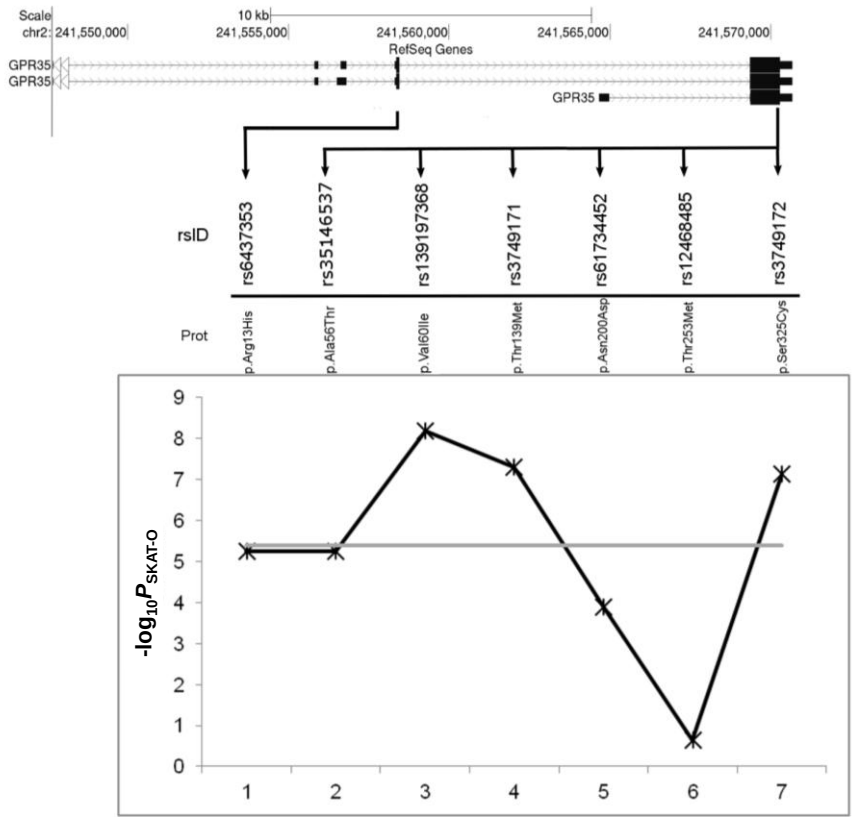


Figure R23. Contribution of individual *GPR35* variants on statistical significances for the *GPR35* gene. Top: genomic location of *GPR35* displayed in the UCSC Genome Browser. Exon location and amino acid substitution of each of the 7 coding polymorphic variants covered by the Illumina HumanExome BeadChip array are depicted. Bottom: *P*-values for the *GPR35* association in SKAT-O gene-based tests after removing one variant (black line) at a time and recalculating the association for *GPR35*. Grey line indicates the *P*-value for the *GPR35* association with chronic AIC including all 7 coding variants (P=7.0x10^{-6}). Abbreviations: Prot, protein.

The minor T allele of this variant was almost exclusively present in cases (MAF_{CASES}=0.09 v $MAF_{CONTROLS}$=0.009), with only one CT carrier (2%) among controls compared to 6 CT carriers (19%) in cases. No TT carriers were found in our series of anthracycline-treated pediatric cancer patients (**Table R11**). The majority (67%) of cases carrying the CT genotype had an extreme chronic AIC phenotype: LV dysfunction, mostly symptomatic, evidenced after treatment with anthracycline doses well below the average for cases (CT cases=155mg/m^2 *v* all other cases=360 mg/m^2). To assess whether the model with the variant rs12468485 was more informative than the model with only non-genetic variables (age at diagnosis, cumulative anthracycline dose and bleomycin concomitant therapy) we used likelihood ratio tests. We

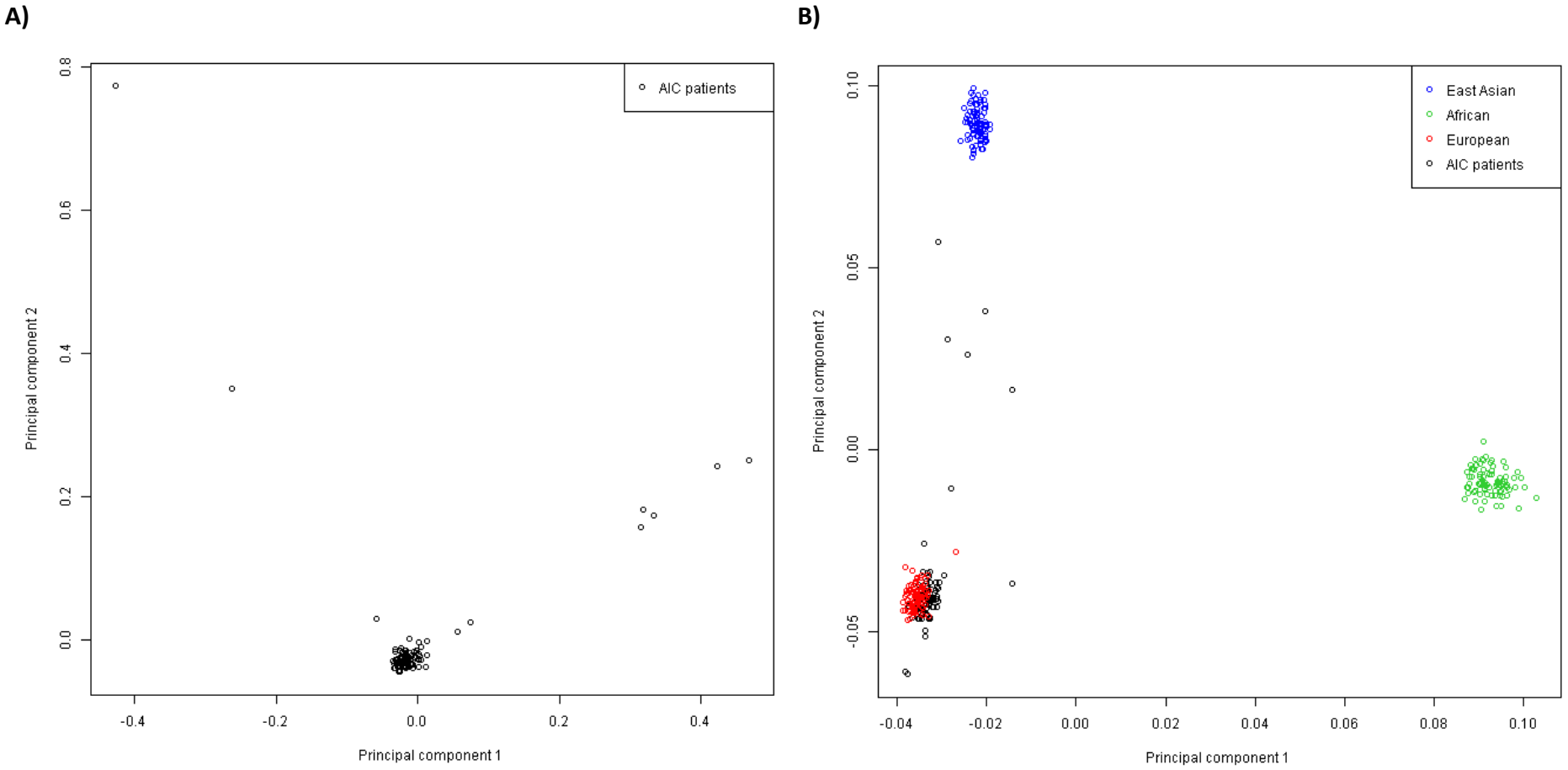


Figure R21. Principal component analysis of genetic data. We carried out PCA analysis using the 90 anthracycline-treated pediatric samples that passed quality control (**A**) and both 90 anthracycline-treated pediatric samples and 277 HapMap samples (European, African and East Asian samples) (**B**) genotyped on the Illumina HumanExome-12v1_A Beadchip by Illumina. We plotted principal component 1 and principal component 2.

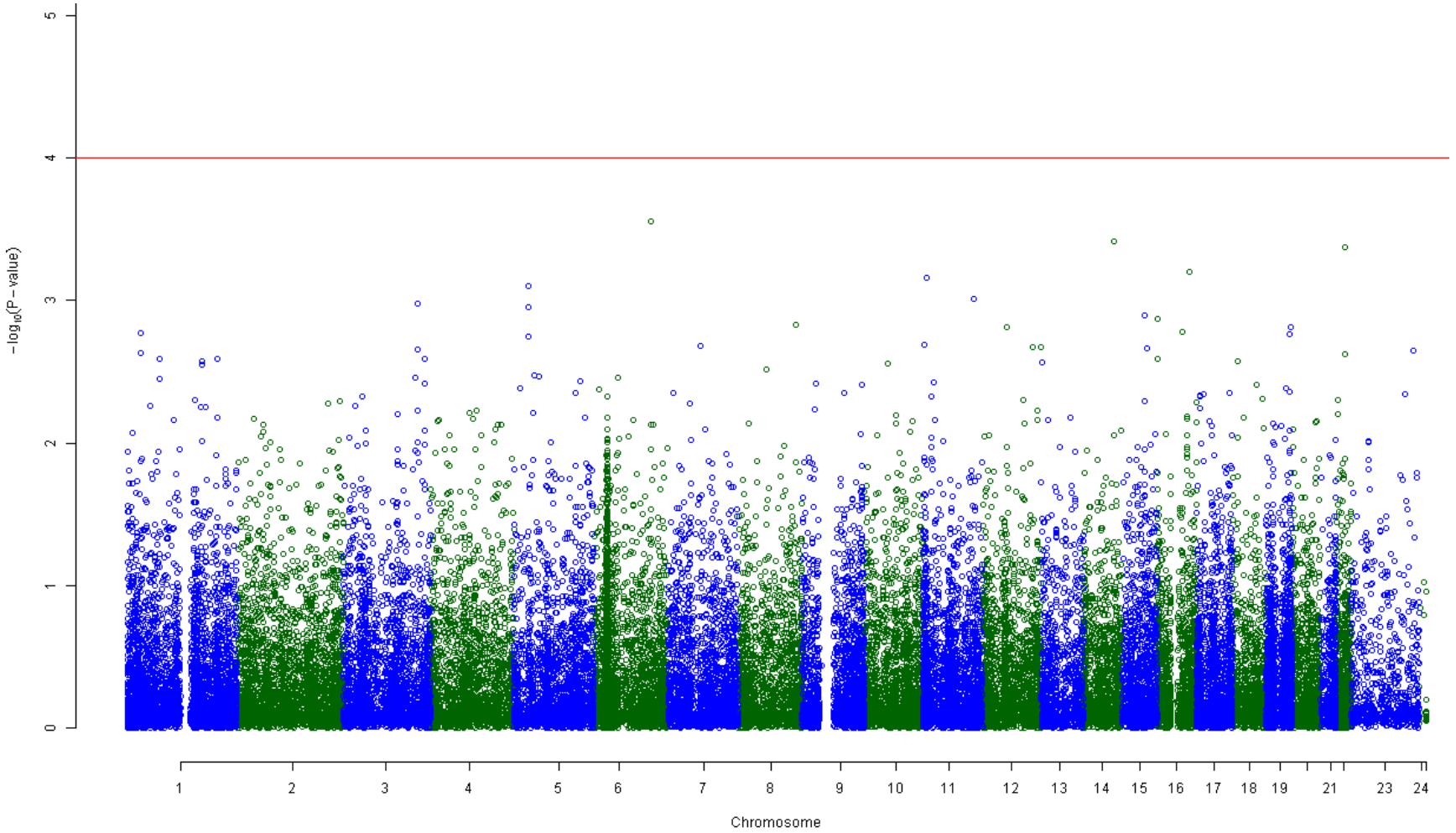


Figure R22. Manhattan plot showing association between the genotypes of the 53,136 polymorphic variants and risk of chronic AIC. The $-\log_{10}$(*P*-values) from single-variant analysis by logistic regression adjusted for age at diagnosis, cumulative anthracycline dose and bleomycin concomitant administration is plotted against its physical chromosomal position. The most significantly associated variant was rs858345 in chromosome 6 (P=2.79 x10^{-4}). Red line: P=10^{-4}.

Table R11. Allelic distribution of the 7 polymorphic *GPR35* variants covered by the Illumina HumanExome BeadChip array

Variant	Function	Chr.	Location*	Genotype	Cases (N=31) N/%	Controls (N=52) N/%	MAF cases	MAF controls	MAF
rs6437353 (A>G)	Missense (p.Arg13His)	2	241558397	AA AG GG	9 (29%) 16 (52%) 6 (19%)	11 (21%) 29 (56%) 12 (53%)	0.45	0.51	0.49
rs35146537 (G>A)	Missense (p.Ala56Thr)	2	241569442	GG GA AA	30 (97%) 1 (3%) -	52 (100%) - -	0.02	-	0.006
rs139197368 (G>A)	Missense (p.Val60Ile)	2	241569454	GG GA AA	31 (100%) - -	51 (98%) 1 (2%) -	-	0.01	0.006
rs3749171 (C>T)	Missense (p.Thr139Met)	2	241569692	CC CT TT	26 (84%) 4 (13%) 1 (3%)	35 (67%) 15 (29%) 2 (4%)	0.1	0.18	0.15
rs61734452 (A>G)	Missense (p.Asn200Asp)	2	241569874	AA AG GG	30 (97%) 1 (3%) -	51 (98%) 1 (2%) -	-	0.01	0.006
rs12468485 (C>T)	Missense (p.Thr253Met)	2	241570127	CC CT TT	25 (81%) 6 (19%) -	51 (98%) 1 (2%) -	0.09	0.009	0.04
rs3749172 (A>C)	Missense (p.Ser325Cys)	2	241570249	AA AC CC	8 (26%) 14 (45%) 9 (29%)	16 (31%) 27 (52%) 9 (17%)	0.52	0.43	0.46

*Chromosome positions are based on Genome Reference Consortium Human Build 37 (GRCh37/hg19). Variant rs12468485 is shown in bold.

obtained that the model including the variant rs12468485 and clinical factors provides a significant improvement over the model with only clinical variables (P=8.6x10^{-5}).

In order to evaluate the impact of the missense variant rs12468485 (p.Thr253Met) on *GPR35* protein structure or function, we applied six in silico prediction algorithms (**Table R12**). p.Thr253Met was classified as pathogenic by PolyPhen-2 and MutPred, and was predicted by F-SNPs to a have a potentially regulatory role in splicing (**Table R12**).

3.3. GPR35 sequencing

Due to the incomplete coverage in the exome array of coding variants in *GPR35*, we sequenced the exonic region of the gene in our series of anthracycline-treated pediatric cancer patients. We identified 17 coding variants, 6 of which had been genotyped on the exome array. Of the other 11, 2 were in complete LD with the variant rs12468485 and 2 had call rate<0.90 and were not analysed further. Of the remaining 7 coding variants identified (r^2<0.36 with variant rs12468485) in *GPR35* (**Table R13**), 4 were synonymous and 3 missense, and all had MAF<5% (MAF ranging from 0.7%-2%).

Table R13. Additional *GPR35* coding variants identified by sequencing

Variant	Position	Function	MAF
rs138283952	chr2:241569810	Synonymous	0.007
rs142918765	chr2:241570284	Synonymous	0.007
rs147336244	chr2:241569669	Synonymous	0.006
rs34778053	chr2:241569742	Missense (p.Arg156Ser)	0.01
rs35155396	chr2:241569585	Synonymous	0.02
rs61734453	chr2:241569745	Missense (p.Gly157Arg)	0.006
rs763867971	chr2:241570132	Missense (p.Arg286Cys)	0.007

Chromosome positions are based on Genome Reference Consortium Human Build 37 (GRCh37/hg19)

Of these 7 new *GPR35* variants, only rs35155396 was associated with risk independently of clinical factors (P=5.16x10^{-3}), but not independently of rs12468485 (P=0.99).

Table R12. In silico prediction of the functional effect of rs12468485 (p.Thr253Met)

Variant	SIFT prediction	Polyphen-2 prediction	MutPred prediction	SNPs&GO Prediction	PON-P2 prediction	PredictSNP prediction	F-SNP prediction			
							ESEfinder prediction	ESRSearch prediction	PESX prediction	RESCUE_ESE prediction
rs12468485 (p.Thr253Met)	Tolerated	**Possibly damaging**	**Pathogenic**	Neutral	Neutral	Neutral	**Changed**	**Changed**	**Changed**	**Changed**

Bold type indicates a likely pathogenic effect or a change in splicing predicted by each in silico algorithm. Predictions are on GPR35 protein with Uniprot identifier Q9HC97.

3.4. Gene-enrichment and pathway analysis

Finally, to gain further insight into the nature of the biological pathways impacting on AIC we performed a gene enrichment analysis using the bioinformatics tool DAVID[168] based on the list of significant genes (P<0.05) with at least 3 variants identified in the SKAT-O analysis. Ten clusters with an enrichment score (ES) ≥1.3 (indicating biological significance) were found, but only 2 clusters had categories with significant P-values after FDR correction (**Table R14**). These clusters revealed overrepresentation of glycoproteins, receptors, N-linked glycosylation sites and plasma membrane components.

On the other hand, no pathways were associated with AIC risk after multiple testing correction (P≥0.05).

Main results Study III

We identified a novel significant association for *GPR35* (G protein-coupled receptor 35) by gene-based testing, a gene with potential roles in cardiac physiology and pathology (P=7.0x10^{-6}), which remained statistically significant after correction for multiple testing (P_{FDR}=0.03). The greatest contribution to this observed association was made by rs12468485, a missense variant (p.Thr253Met, c.758C>T, MAF=0.04), the T allele being associated with increased risk of chronic AIC and more severe symptomatic cardiac manifestations at low anthracycline doses in pediatric cancer patients. No single-variant showed an association with chronic AIC.

Table R14. Functional Annotation Clustering from the DAVID tool (Enrichment score ≥ 1.3).

Annotation cluster 1	**Enrichment score: 3.1**				
Category	**Term**	**Count**	***P***	**Genes**	**P_{FDR}**
SP_PIR_KEYWORDS	Glycoprotein	125	$9.2x10^{-7}$	*FSTL4, OR8S1, GRIN3B, SLC7A4, MMRN2, LPHN2, OR4D2, CHRNA9, CD44, LRRC52, PI16, ODZ3, GRID1, CLCA1, OR10S1, SLCO4A1, SPARCL1, PTPRN2, PLXNB2, CDHR2, SLC22A20, HLA-C, PNPLA3, TNFAIP6, LRP11, OR8B4, CHGB, DST, FGFR4, PANX1, AMTN, OR2T1, KEL, ENPP3, ITGAM, CRB2, P2RY2, ENTPD7, BAI2, SFTPD, B3GNT3, EGF, GCNT1, ADAM28, EPB41, ITGA3, OR5AC2, GPR35, SLC4A11, P2RX3, NOTCH4, TSC2, OR51A7, UTP14C, COL20A1, CHRNG, ARSB, ARSD, CLSTN2, CLSTN3, TMEM161A, CDSN, FCRL3, RSPO4, C14ORF135, SMPDL3A, NMUR2, POMT1, OR6P1, ANO2, CTBS, USH2A, OR2AE1, OR10J1, NCR3, CD86, TNFRSF10C, CLECL1, TAS2R19, ZPBP2, GRM6, ERN1, KCNH8, C4ORF29, PLA2G3, WFIKKN2, LRRC32, CPN2, NUP214, C1QTNF6, FCN2, COL6A2, FUT3, UGT2A1, COL6A1, CCBP2, HRG, PLA2R1, DKKL1, SLC39A4, QSOX1, SOAT1, TMC6, MOGAT1, CES1, OR8G2, FBN1, FUCA2, MAN1C1, COL5A1, PLG, OR51M1, CDH13, PROM2, GPR110, SLC17A2, ANXA11, GPR113, MEP1B, GPR111, MEGF6, ABCC8, GFRA2, ABCC6, GPR116*	**$1.7x10^{-4}$**
UP_SEQ_FEATURE	N-linked glycosylation sites (GlcNAc...)	116	$1.7x10^{-5}$	*FSTL4, OR8S1, GRIN3B, SLC7A4, MMRN2, LPHN2, OR4D2, CHRNA9, CD44, LRRC52, PI16, ODZ3, GRID1, CLCA1, OR10S1, SPARCL1, SLCO4A1, PTPRN2, PLXNB2, CDHR2, SLC22A20, HLA-C, PNPLA3, TNFAIP6, LRP11, OR8B4, DST, FGFR4, PANX1, OR2T1, KEL, ENPP3, ITGAM, CRB2, P2RY2, ENTPD7, BAI2, SFTPD, B3GNT3, EGF, GCNT1, ADAM28, ITGA3, OR5AC2, GPR35, SLC4A11, P2RX3, NOTCH4, OR51A7, UTP14C, COL20A1, CHRNG, ARSB, ARSD, CLSTN2, CLSTN3, TMEM161A, CDSN, FCRL3, RSPO4, C14ORF135, SMPDL3A, NMUR2, POMT1, OR6P1, ANO2, CTBS, USH2A, OR2AE1, OR10J1, NCR3, CD86, CLECL1, TAS2R19, ZPBP2, GRM6, ERN1, KCNH8, C4ORF29, PLA2G3, WFIKKN2, LRRC32, CPN2, C1QTNF6, FCN2, COL6A2, FUT3, UGT2A1, COL6A1, CCBP2, HRG, PLA2R1, DKKL1, SLC39A4, QSOX1, TMC6, MOGAT1, CES1, OR8G2, FBN1, FUCA2, MAN1C1, PLG, OR51M1, CDH13, PROM2, GPR110, SLC17A2, GPR113, MEP1B, GPR111, MEGF6, ABCC8, GFRA2, ABCC6, GPR116*	**$1.2x10^{-2}$**

Continue on next page

Continued

Category	Term	Count	*P*	Genes	P_{FDR}
SP_PIR_KEYWORDS	Receptor	55	$3.3x10^{-5}$	*TRPV3, GRIN3B, OR8S1, FCRL3, LPHN2, OR4D2, CHRNA9, CD44, NMUR2, OR6P1, GRID1, OR10S1, PLXNB2, PTPRN2, PPARGC1A, OR2AE1, OR10J1, TRPM2, NCR3, TNFRSF10C, CD86, TAS2R19, LILRB4, GRM6, LRP11, GPR50, OR8B4, FGFR4, OR2T1, PAQR7, ITGAM, P2RY2, BAI2, CCBP2, PLA2R1, TRIP11, OR8G2, ITGA3, PTPN12, OR5AC2, OR51M1, GPR35, EPS8, GPR110, P2RX3, NOTCH4, GPR113, GPR111, MS4A10, ABL1, ABCC8, OR51A7, GFRA2, CHRNG, GPR116*	**$2.5x10^{-3}$**
GOTERM_CC_FAT	Plasma membrane	107	$5.3x10^{-5}$	*LMO7, OR8S1, GRIN3B, LPHN2, OR4D2, CHRNA9, CD44, WNK4, GRID1, CLCA1, OR10S1, MYH1, PTPRN2, CDHR2, HLA-C, CTNNA3, LILRB4, GPR50, OR8B4, KBTBD10, DST, SNTG2, HDLBP, FGFR4, PANX1, AMTN, OR2T1, KEL, ENPP3, MAP4K2, PAQR7, ITGAM, KCNS3, CRB2, PLCH2, P2RY2, TRO, BAI2, B3GNT3, EGF, SLC28A3, ADAM28, EPB41, LPP, KCNB1, ITGA3, OR5AC2, GPR35, SLC4A11, EPS8, P2RX3, NOTCH4, TSC2, MAP7, OR51A7, PLA2G4E, CHRNG, TM7SF4, CLSTN2, CLSTN3, KCNJ12, CDSN, FCRL3, NMUR2, SDPR, OR6P1, KCNG4, RHOF, USH2A, OR2AE1, TRPM2, OR10J1, NCR3, CD86, TNFRSF10C, CLECL1, SLC26A9, GRM6, PLA2G3, NKD2, LRRC32, TAP2, COL6A2, COL6A1, CCBP2, PLA2R1, SLC39A4, OR8G2, PPP1R9B, COG4, OR51M1, CDH13, PROM2, ERBB2IP, NRAP, GPR110, SLC17A2, GPR113, MEP1B, ANXA13, GPR111, SYNM, TAPBPL, ABCC8, GFRA2, ABCC6, GPR116*	**$1.5x10^{-2}$**

Continue on next page

Annotation cluster 2	**Enrichment score: 2.19**				
Category	**Term**	**Count**	***P***	**Genes**	**P_{FDR}**
SP_PIR_KEYWORDS	Glycoprotein	125	9.2×10^{-7}	*FSTL4, OR8S1, GRIN3B, SLC7A4, MMRN2, LPHN2, OR4D2, CHRNA9, CD44, LRRC52, PI16, ODZ3, GRID1, CLCA1, OR10S1, SLCO4A1, SPARCL1, PTPRN2, PLXNB2, CDHR2, SLC22A20, HLA-C, PNPLA3, TNFAIP6, LRP11, OR8B4, CHGB, DST, FGFR4, PANX1, AMTN, OR2T1, KEL, ENPP3, ITGAM, CRB2, P2RY2, ENTPD7, BAI2, SFTPD, B3GNT3, EGF, GCNT1, ADAM28, EPB41, ITGA3, OR5AC2, GPR35, SLC4A11, P2RX3, NOTCH4, TSC2, OR51A7, UTP14C, COL20A1, CHRNG, ARSB, ARSD, CLSTN2, CLSTN3, TMEM161A, CDSN, FCRL3, RSPO4, C14ORF135, SMPDL3A, NMUR2, POMT1, OR6P1, ANO2, CTBS, USH2A, OR2AE1, OR10J1, NCR3, CD86, TNFRSF10C, CLECL1, TAS2R19, ZPBP2, GRM6, ERN1, KCNH8, C4ORF29, PLA2G3, WFIKKN2, LRRC32, CPN2, NUP214, C1QTNF6, FCN2, COL6A2, FUT3, UGT2A1, COL6A1, CCBP2, HRG, PLA2R1, DKKL1, SLC39A4, QSOX1, SOAT1, TMC6, MOGAT1, CES1, OR8G2, FBN1, FUCA2, MAN1C1, COL5A1, PLG, OR51M1, CDH13, PROM2, GPR110, SLC17A2, ANXA11, GPR113, MEP1B, GPR111, MEGF6, ABCC8, GFRA2, ABCC6, GPR116*	SP_PIR_KEYWORDS
UP_SEQ_FEATURE	N-linked glycosylation sites (GlcNAc...)	116	1.7×10^{-5}	*FSTL4, OR8S1, GRIN3B, SLC7A4, MMRN2, LPHN2, OR4D2, CHRNA9, CD44, LRRC52, PI16, ODZ3, GRID1, CLCA1, OR10S1, SPARCL1, SLCO4A1, PTPRN2, PLXNB2, CDHR2, SLC22A20, HLA-C, PNPLA3, TNFAIP6, LRP11, OR8B4, DST, FGFR4, PANX1, OR2T1, KEL, ENPP3, ITGAM, CRB2, P2RY2, ENTPD7, BAI2, SFTPD, B3GNT3, EGF, GCNT1, ADAM28, ITGA3, OR5AC2, GPR35, SLC4A11, P2RX3, NOTCH4, OR51A7, UTP14C, COL20A1, CHRNG, ARSB, ARSD, CLSTN2, CLSTN3, TMEM161A, CDSN, FCRL3, RSPO4, C14ORF135, SMPDL3A, NMUR2, POMT1, OR6P1, ANO2, CTBS, USH2A, OR2AE1, OR10J1, NCR3, CD86, CLECL1, TAS2R19, ZPBP2, GRM6, ERN1, KCNH8, C4ORF29, PLA2G3, WFIKKN2, LRRC32, CPN2, C1QTNF6, FCN2, COL6A2, FUT3, UGT2A1, COL6A1, CCBP2, HRG, PLA2R1, DKKL1, SLC39A4, QSOX1, TMC6, MOGAT1, CES1, OR8G2, FBN1, FUCA2, MAN1C1, PLG, OR51M1, CDH13, PROM2, GPR110, SLC17A2, GPR113, MEP1B, GPR111, MEGF6, ABCC8, GFRA2, ABCC6, GPR116*	UP_SEQ_FEATURE

Gene-set enrichment analysis of gene-based *P*-values from SKAT-O was performed using the functional annotation clustering analysis module of the bioinformatic tool DAVID. Each annotation term group is assigned an enrichment score to rank overall importance. Only annotation clusters with ES≥1.3 (indicating biological significance) and significant *P*-values after FDR correction are shown. The significance of gene-term enrichment was assessed with a modified Fisher´s exact test and *P*-values are corrected using Benjamini-Hochberg's by false discovery rate (FDR-BH) procedure.

4. Results, Study IV: identification of genetic variants predictive of susceptibility to chronic anthracycline-induced cardiotoxicity (AIC) in breast and pediatric oncology patients

The demographic and clinical characteristics of the Spanish and Belgium adult breast cancer patients are shown in **Table R15.**

Table R15. Clinical characteristics of the anthracycline-treated breast cancer patients

	Spanish breast cancer patients (*N*=71)				Belgium breast cancer patients (*N*=142)			
	Controls (*N*=53)		Cases (*N*=18)		Controls (*N*=86)		Cases (*N*=56)	
Characteristic	**N**	**%***	**N**	**%***	**N**	**%***	**N**	**%***
Age at diagnosis (years)								
Median	49		59.5		52		49	
Range	27-73		36-72		32-70		34-74	
Primary diagnosis (tumor type)								
Ductal	42	79	13	72	59	69	42	75
Lobular	8	15	4	22	7	8.1	4	7.1
Others	3	5.7	1	5.6	20	23	10	18
Tumor grade								
1	1	1.9	-	-	6	7	3	5.4
2	36	68	14	78	29	34	19	34
3	16	30	4	22	51	59	34	61
Left-sided radiotherapy	25	47	6	33	49	60	27	48
Anthracycline type								
Doxorubicin	53	100	18	100	-	-	-	-
Epirubicin	-	-	-	-	86	100	56	100
Anthracycline setting								
Neoadjuvant	25	47	13	72	71	83	4	7.1
Adjuvant	28	53	5	28	15	17	52	93
Cumulative anthracycline dose								
Median (mg/m^2)	298.6		298.4		300		300	
Mean (mg/m^2)	282.9		298.1		353.5		428.6	
Range (mg/m^2)	150-375		200-588		300-600		300-600	
LVEF at baseline (%)								
Median	68		75		65		60	
Range	60-83		55-81		60-81		60-86	
LVEF at follow-up (%)								
Median	65		59.5		60		15	
Range	60-82		24-70		60-81		10-58	
Follow-up (years)								
Median	4.76		5.74		7.1		8.04	
Range	2-16		1.19-10.07		1.38-14.6		1.94-13.43	

* Percentages are computed based on the total number of non-missing values.

In the Spanish breast cancer cohort, 18 patients (25%) developed AIC, whereas in the Belgium cohort, 56 patients (39%) experienced cardiotoxicity. Controls were significantly younger than cases at diagnosis in the discovery (median age 49 *v* 59.5 years, respectively; *P*=0.023), but

patients have similar age in the Belgium (52 *v* 49 years, respectively; *P*=0.24). Patients in the Spanish cohort were treated with doxorubicin, while Belgium breast cancer patients were treated with epirubicin. Cumulative anthracycline doses were significantly different between cases and controls in the Belgium cohort (*P*=0.001), with 82% of controls receiving 300 mg/m^2 of epirrubicin and with a median cumulative dose of 353.5 mg/m^2, compared to a median cumulative dose of 428.6 mg/m^2 in cases (54% of cases received 300 mg/m^2 of epirrubicin and 41% of cases received 600 mg/m^2).

In the Spanish breast cancer cohort 1 case was failed heterozigosity and 9 patients (2 cases and 7 controls) were excluded as ethnic outliers based on inspection of plots of the two first principal components, leaving 61 patients for further analysis (**Figure R24**).

4.1. Single-variant associations

Of the 247,870 variants on the array, 246,060 passed quality control and 48,451 were polymorphic. Age at diagnosis was included as covariate for logistic regression analyses. The strongest evidence of association was found for the variant rs1243647 located in the *POLE* (DNA polymerase epsilon) gene at chromosome 14 (*P*=8.6x10^{-4}, OR=15.1) (**Figure R25**). As could be expected and as in the single-variant analyses performed in the pediatric cohort (**Study III**), given the small number of patients relative to the number of genetic variants tested, this variant did not survive a correction for multiple comparisons (P_{FDR}=0.96).

4.2. Gene-based associations

We then carried out gene-based tests in the Spanish breast cancer cohort by using SKAT-O[55,166] to further investigate the join effects of variants within each gene and considering age at diagnosis, as covariate. In total, we tested 4,288 genes of 17,677 covered by at least one variant on the array.

The two genes most significantly associated with chronic AIC in the Spanish breast cancer patients were *ETFB* (electron transfer flavoprotein beta subunit), a cardiac protein involved in mitochondrial energy production[192] (*P*=4.2x10^{-4}) and *WISP1* (*WNT1* inducible signaling pathway protein 1), which has been shown to inhibit the doxorubicin-induced cardiomyocyte death[193] (*P*=5.2x10^{-4}); although they did not remain statistically significant after correction for multiple testing (P_{FDR}=0.77). Given that these genes are, *a priori*, good candidates for influencing the development of chronic AIC and the SKAT-O does not provide any parameter estimates; we

assessed the individual contribution of variants within *ETFB* and *WISP1*. Variant rs79338777 (p.Pro52Leu; c.155C>T) in *ETFB* (**Figure R26**) and variants rs149172980 (p.Thr13Ile; c.38C>T), rs72731540 (p.Val184Ile; c.550G>A) and rs143089011 (p.Ala196Thr; c.586G>A) in *WISP1* (**Figure R27**) made the greatest contribution to the observed association. The minor T allele of the *ETFB* variant rs79338777 variant was almost exclusively present in cases (MAF_{CASES}=13% *v* $MAF_{CONTROLS}$=2%), with 4 CT carriers (27%) among cases compared to 2 CT carriers (4%) in controls (**Table R16**). *WISP1* variants rs149172980 and rs143089011 were found in heterozygosis in only 1 case (3% of cases) out of the 61 Spanish breast cancer patients (MAF_{CASES}=3%) with no patients carying the minor allele in homozygosis. The minor A allele of the *WISP1* rs72731540 variant was found only in cases, with only 2 GA carriers (13%) (MAF_{CASES}=7%) (**Table R16**).

Next we decided to evaluate the distribution of the allele frequencies of these 4 low-frequency variants in two independent cohorts of anthracycline-treated cancer patients: in 142 Belgium breast cancer patients and in the Spanish pediatric cancer seires used in **Study III**. As in the Spanish breast cancer cohort, the minor T allele of the variant rs79338777 (*ETFB*) was found more often in Belgium breast cancer cases than in controls (9 CT_{CASES} (16%) *v* 6 $CT_{CONTROLS}$ (7%); MAF_{CASES}=8% *v* $MAF_{CONTROLS}$=4%). Regarding *WISP1* variants, the minor A allele of the rs72731540 variant was consistently and mostly found in cases in the Belgium cohort controls (4 GA_{CASES} (7%) *v* 2 $CT_{CONTROLS}$ (2%); MAF_{CASES}=4% *v* $MAF_{CONTROLS}$=1%) (**Table R17**). Unfortunately, for the other two *WISP1* variants, we found that rs149172980 and rs143089011 were monomorphic in the Belgium cohort.

Consistently with what we found in the Spanish and Belgium breast cancer series, the minor allele of variants rs79338777 (*ETFB*) and rs72731540 (*WISP1*) was more common in cases than in pediatric controls (rs79338777, 9 CT_{CASES} (29%) *v* 7 $CT_{CONTROLS}$ (13%), MAF_{CASES}=18% *v* $MAF_{CONTROLS}$=7%; rs72731540, 2 GA_{CASES} (6%) *v* 2 $GA_{CONTROLS}$ (4%), MAF_{CASES}=3% *v* $MAF_{CONTROLS}$=2%) (**Table R17**). Variants rs149172980 and rs143089011 in *WISP1* were also monomorphic in children.

When we combined the three cohorts to show overall significance levels (*N*=286), we obtained that variants rs79338777 (*ETFB*) and rs72731540 (*WISP1*) were significantly associated with risk of chronic AIC [rs79338777 (*ETFB*), OR=3.55, $P=7.71\times10^{-4}$, 95%CI=1.70–7.42; rs72731540 (*WISP1*), OR=5.85, *P*=0.0084, 95%CI=1.57–21.7] (**Table R17**). Overall, it was found that the risk allele for rs79338777 (*ETFB*) to be 2.89 times more frequent in cases than in

controls (MAF_{CASES}=12% *v* $MAF_{CONTROLS}$=4%) and 3.61 times (MAF_{CASES}=4% *v* $MAF_{CONTROLS}$=1%) for rs72731540 (*WISP1*).

In order to evaluate the impact of the low-frequency missense variants rs79338777 (p.Pro52Leu) and rs72731540 (p.Val184Ile) on ETFB and WISP1 protein structure or function, respectively, we applied six in silico prediction algorithms. rs79338777 (*ETFB*) was classified as pathogenic by Predict-SNP, and SIFT and as possibly damaging by PolyPhen-2; whereas rs72731540 (*WISP1*) was classified as non-pathogenic by all the prediction methods (**Table R18**).

4.3. Gene-enrichment and pathway analysis

Finally, to gain further insight into the nature of the biological pathways impacting on AIC we performed a gene enrichment analysis using the bioinformatics tool DAVID[168] based on the list of significant genes (P<0.05) with at least 3 variants identified in the SKAT-O analysis. 2 clusters with an enrichment score (ES) ≥1.3 (indicating biological significance) with significant *P*-values after FDR correction were found (**Table R19**). These clusters revealed overrepresentation of glycoproteins, N-linked glycosylation sites, cytoplasmic and extracellular components.

On the other hand, no pathways were associated with AIC risk after multiple testing correction (P≥0.05).

Main results Study IV

By gene-based testing we identified novel associations for two genes, *ETFB* (electron transfer flavoprotein beta subunit), a cardiac protein involved in mitochondrial energy production (P=4.2x10^{-4}) and *WISP1* (WNT1 inducible signaling pathway protein 1), which has been shown to inhibit doxorubicin-induced cardiomyocyte death (P=5.2x10^{-4}). The low-frequency alleles of *ETFB* variant rs79338777 (p.Pro52Leu; c.155C>T) and *WISP1* rs72731540 (p.Val184Ile; c.550G>A) variant were associated with increased risk of chronic AIC in the three cohorts when analyzed separately and when combined (rs79338777, OR=3.55, P=7.71×10^{-4}, 95%CI=1.70–7.42; rs72731540, OR=5.85, P=0.0084, 95%CI=1.57–21.7).

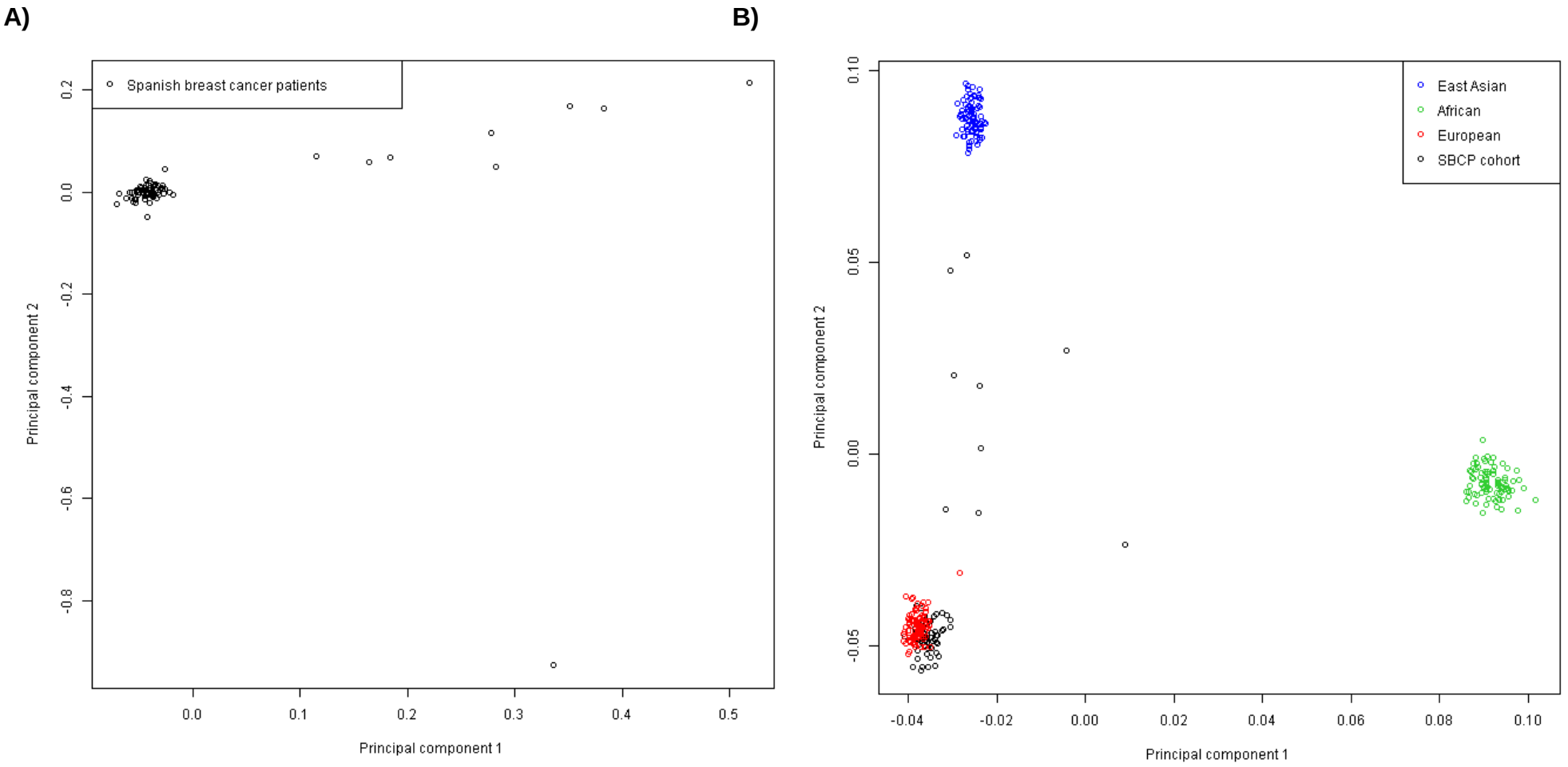


Figure R24. Principal component analysis of genetic data. We carried out PCA analysis using the 70 anthracycline-treated Spanish breast cancer samples that passed quality control (**A**) and both 70 anthracycline-treated breast cancer samples and 277 HapMap samples (European, African and East Asian samples) (**B**) genotyped on the Illumina HumanExome-12v1_A Beadchip by Illumina. We plotted principal component 1 and principal component 2. Abbreviations: SBCP, Spanish breast cancer patients.

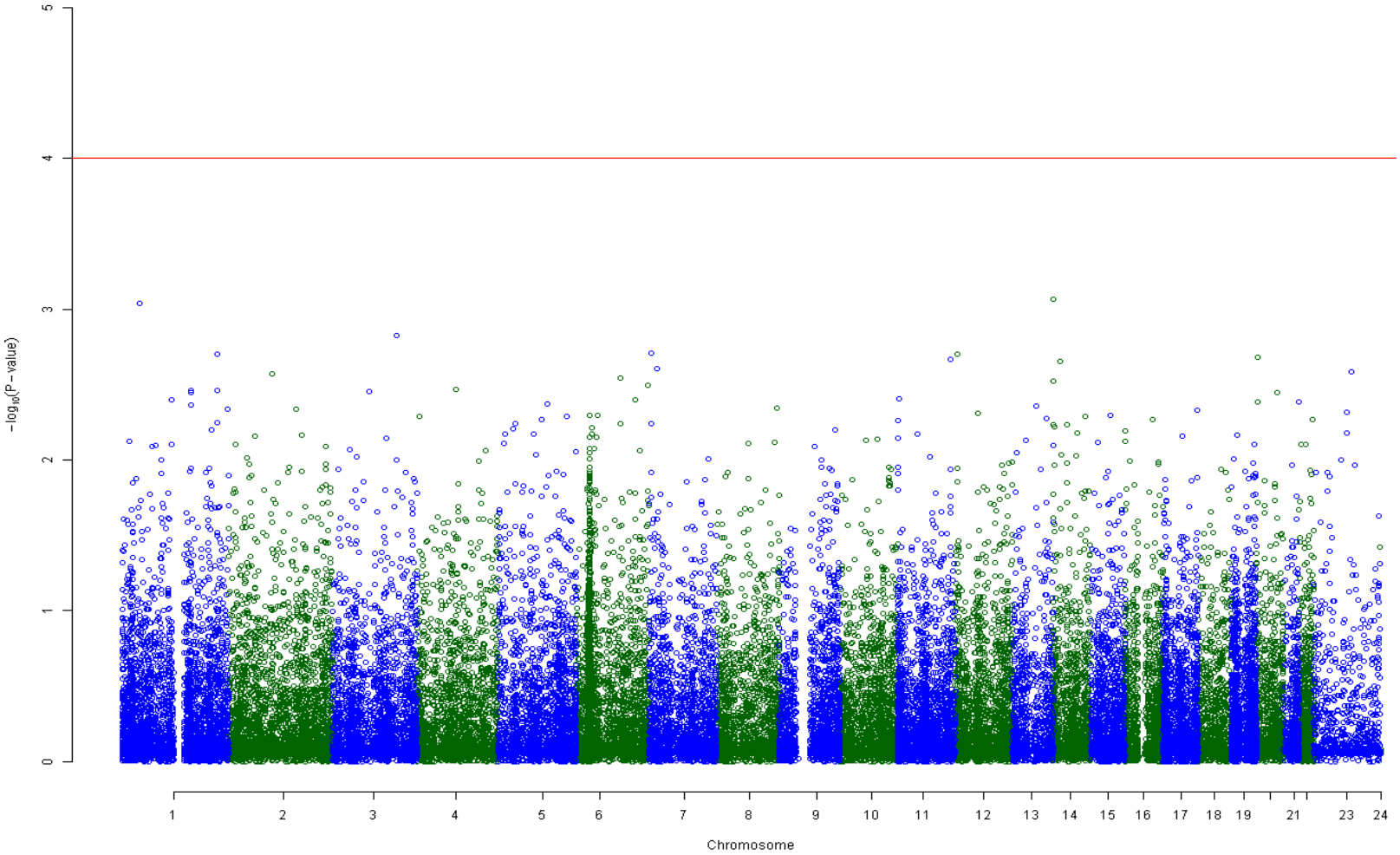


Figure R25. Manhattan plot showing association between the genotypes of the 48,451 polymorphic variants and risk of chronic AIC in the Spanish breast cancer patient cohort. The $-\log_{10}$(*P*-values) from single-variant analysis by logistic regression adjusted for age at diagnosis is plotted against its physical chromosomal position. The most significantly associated variant was rs1243647 in chromosome 14 (P=8.6 x10^{-4}). Red line: P=10^{-4}.

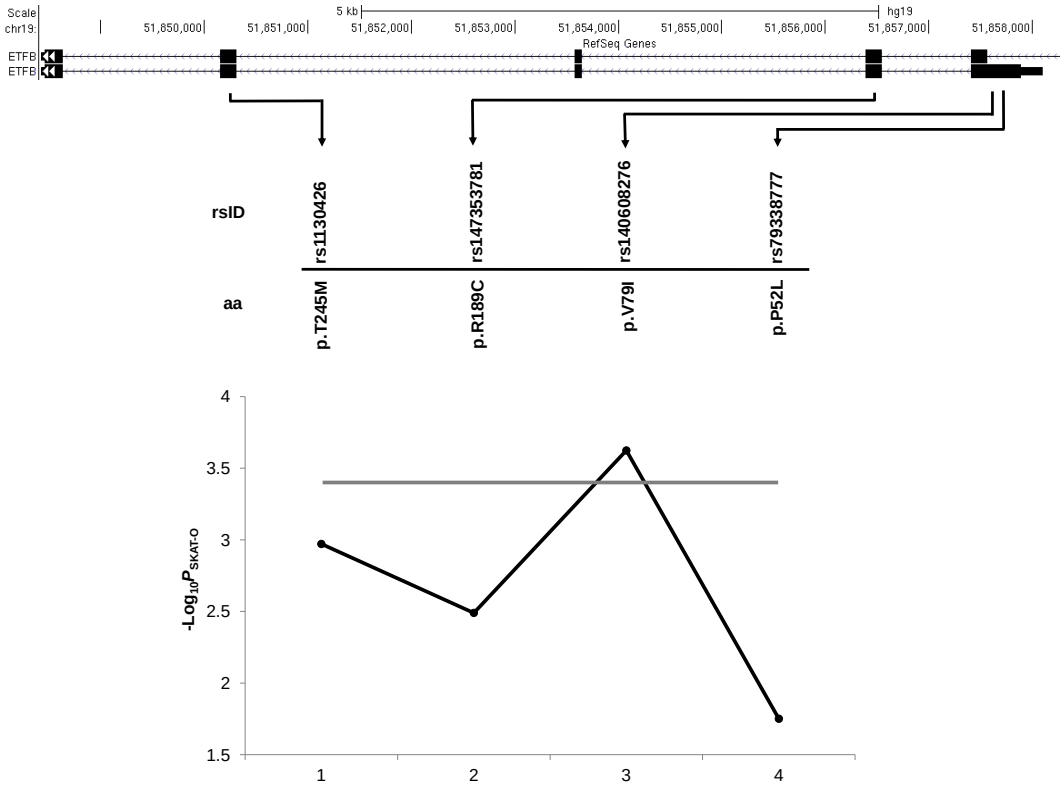


Figure R26. Contribution of individual *ETFB* variants on statistical significances for the *ETFB* gene. Top: genomic location of *ETFB* displayed in the UCSC Genome Browser. Exon location and amino acid substitution of each of the 4 coding polymorphic variants covered by the Illumina HumanExome BeadChip array are depicted. Bottom: *P*-values for the *ETFB* association in SKAT-O gene-based tests after removing one variant (black line) at a time and recalculating the association for *ETFB*. Grey line indicates the *P*-value for the *ETFB* association with chronic AIC including all 4 coding variants (P=4.2x10^{-4}).

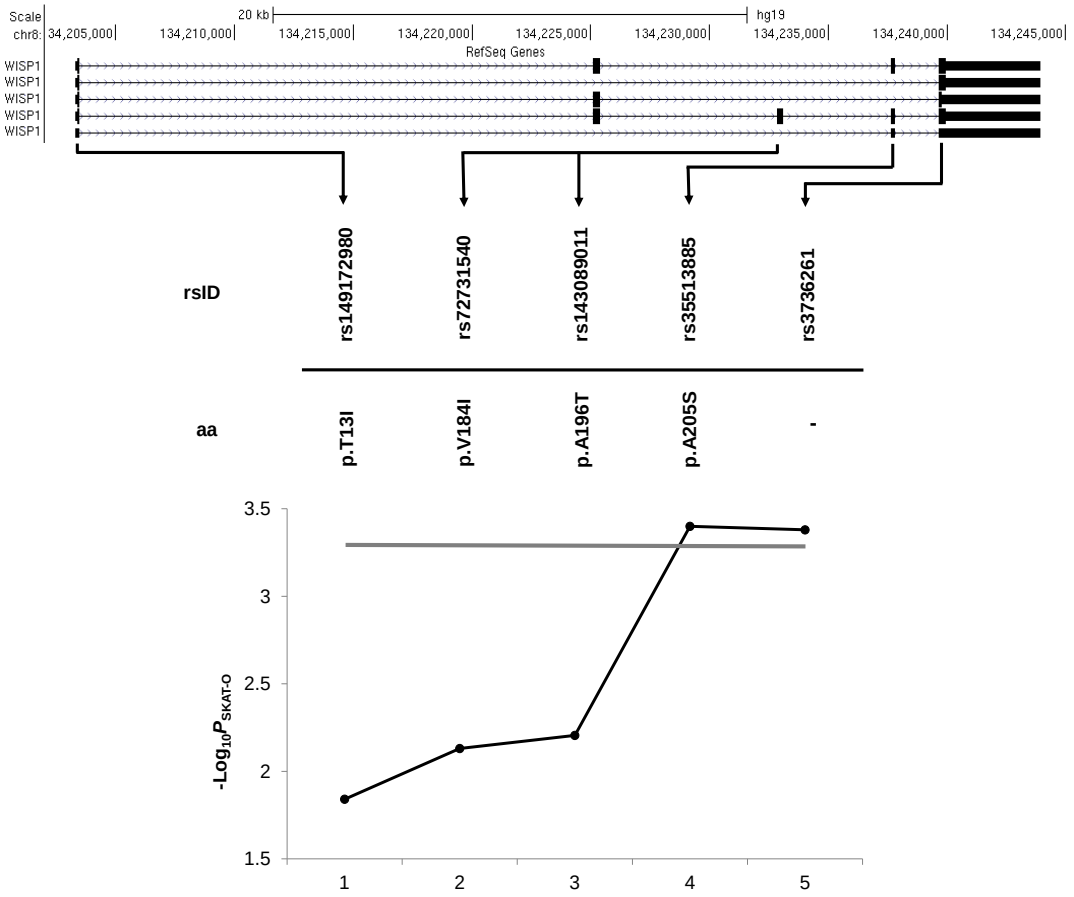


Figure R27. Contribution of individual *WISP1* variants on statistical significances for the *WISP1* gene. Top: genomic location of *WISP1* displayed in the UCSC Genome Browser. Exon location and amino acid substitution of each of the 5 coding polymorphic variants covered by the Illumina HumanExome BeadChip array are depicted. Bottom: *P*-values for the *ETFB* association in SKAT-O gene-based tests after removing one variant (black line) at a time and recalculating the association for *WISP1*. Grey line indicates the *P*-value for the *WISP1* association with chronic AIC including all 5 coding variants (P=5.2x10^{-4}). rs3736261 is a synonymous variant.

Table R16. Allelic distribution of the polymorphic *ETFB* and *WISP1* variants covered by the Illumina HumanExome BeadChip array in the Spanish breast cancer cohort

Variant	Gene	Function	Chr.	Location*	Genotype	Cases (N=15) N/%	Controls (N=46) N/%	MAF cases	MAF controls	MAF
rs1130426 (C>T)	*ETFB*	Missense (p.Thr245Met)	19	51850290	CC CT TT	7 (27%) 4 (46%) 7 (27%)	15 (33%) 23 (50%) 8 (17%)	0.5	0.42	0.43
rs147353781 (C>T)	*ETFB*	Missense (p.Arg189Cys)	19	51856469	CC CT TT	14 (97%) 1 (3%) -	46 (100%) - -	0.03	-	0.008
rs140608276 (G>A)	*ETFB*	Missense (p.Val79Ile)	19	51857658	GG GA AA	14 (97%) 1 (3%) -	46 (100%) - -	0.03	-	0.008
rs79338777 (C>T)	*ETFB*	Missense (p.Pro52Leu)	19	51857738	CC CT TT	11 (73%) 4 (27%) -	44 (96%) 2 (4%) -	0.13	0.02	0.04
rs149172980 (C>T)	*WISP1*	Missense (p.Thr13Ile)	8	134203425	CC CT TT	14 (97%) 1 (3%) -	46 (100%) - -	0.03	-	0.008
rs72731540 (G>A)	*WISP1*	Missense (p.Val184Ile)	8	134233024	GG GA AA	13 (87%) 2 (13%) -	46 (100%) - -	0.07	-	0.04
rs143089011 (G>A)	*WISP1*	Missense (p.Ala196Thr)	8	134233060	GG GA AA	14 (97%) 1 (3%) -	46 (100%) - -	0.03	-	0.008
rs35513885 (G>T)	*WISP1*	Missense (p.Ala205Ser)	8	134237635	GG GT TT	13 (87%) 2 (13%) -	40 (87%) 5 (11%) 1 (2%)	0.07	0.08	0.07
rs3739261 (T>C)	*WISP1*	Synonymous	8	134239770	TT TC CC	8 (53%) 4 (27%) 3 (20%)	19 (41%) 23 (50%) 4 (9%)	0.33	0.34	0.34

*Chromosome positions are based on Genome Reference Consortium Human Build 37 (GRCh37/hg19)

Table R17. Association of *ETFB* and *WISP1* variants with anthracycline-induced cardiotoxicity

Variant	Gene	Genotype	Belgium breast cancer patients (N=142)					Spanish pediatric oncology patients (N=83)					Overall combined logistic regression *N*=286			
			Cases (N=56) N/%	Controls (N=86) N/%	MAF cases	MAF controls	MAF	Cases (N=31) N/%	Controls (N=52) N/%	MAF cases	MAF controls	MAF	*P*	OR	95%CI	MAF
rs79338777 (C>T)	*ETFB*	CC	47 (84%)	78 (91%)				21 (68%)	45 (87%)							
		CT	9 (16%)	6 (7%)	0.08	0.04	0.05	9 (29%)	7 (13%)	0.18	0.07	0.11	7.71×10^{-4}	3.55	1.70–7.42	0.07
		TT	-	1 (1%)				1 (3%)	-							
rs72731540 (G>A)	*WISP1*	GG	52 (93%)	83 (97%)				29 (94%)	50 (96%)							
		GA	4 (7%)	2 (2%)	0.04	0.01	0.02	2 (6%)	2 (4%)	0.03	0.02	0.02	0.0084	5.85	1.57–21.7	0.02
		AA	-	-				-	-							

WIPS1 variants rs149172980 and rs143089011 were monomorphic in the Belgium breast cancer and in the Spanish pediatric cohorts

Table R18. In silico prediction of the functional effect of *ETFB* rs79338777 and *WISP1* rs72731540

Variant	SIFT prediction	Polyphen-2 prediction	MutPred prediction	SNPs&GO Prediction	PON-P2 prediction	PredictSNP prediction
rs79338777 (p.Pro52Leu)	**Deleterious**	**Possibly damaging**	Non pathogenic	-	-	**Deleterious**
rs72731540 (p.Val184Ile)	Tolerated	Benign	Neutral	Neutral	Neutral	Neutral

Bold type indicates a likely pathogenic effect. Predictions are on ETFB protein with Uniprot identifier P38117-2 and WISP1 protein with Uniprot identifier O95388. rs79338777 is located in the ETFB isoform 2, which is not available in the SNPs&GO and PON-P2 web servers

Table R19. Functional Annotation Clustering from the DAVID tool (Enrichment score ≥ 1.3).

Annotation cluster 1	**Enrichment score: 2.79**				
Category	**Term**	**Count**	***P***	**Genes**	**P_{FDR}**
UP_SEQ_FEATURE	N-linked glycosylation sites (GlcNAc...)	72	$1.8x10^{-5}$	*PVR, QRFPR, SLC8A3, A2M, OR10A4, SLC44A2, LTBP1, GBGT1, MMP8, FAM20C, CHPF2, RAET1E, WISP1, POMT1, VNN1, CEACAM5, LBP, CHST15, COL11A1, LAG3, CEACAM20, MATN4, KCND1, PCDHB6, OLFML2B, PCDHB4, CILP, BTNL2, PCDHGB6, SLC22A20, PKD2L1, OR6V1, GLRA4, LAMC1, GLP1R, TG, HLA-DQB1, CALCR, GPR108, FGFR4, SLC2A11, SERPINA10, BTN2A1, OR51F2, PSAPL1, CLEC10A, OR2G2, VWA5B1, DISP1, OR6A2, NUP210, UPK1A, BTN1A1, GALNS, EGF, PCSK6, ST6GAL2, CPA6, FOXRED2, OR6C74, STAB2, GPR151, SERPINI2, PXDNL, LAMA4, SLC4A11, STAB1, SLC17A1, P2RX2, DSC2, ADRA1A, ABCC2*	**$4.7x10^{-3}$**
UP_KEYWORDS	Glycoprotein	74	$3.8x10^{-5}$	*PVR, QRFPR, SLC8A3, A2M, OR10A4, SLC44A2, LTBP1, GBGT1, MMP8, FAM20C, CHPF2, RAET1E, WISP1, POMT1, VNN1, CEACAM5, LBP, CHST15, ANO7, COL11A1, LAG3, CEACAM20, MATN4, KCND1, PCDHB6, OLFML2B, PCDHB4, CILP, BTNL2, PCDHGB6, SLC22A20, PKD2L1, OR6V1, GLRA4, LAMC1, GLP1R, TG, HLA-DQB1, CALCR, GPR108, FGFR4, SLC2A11, SERPINA10, BTN2A1, OR51F2, PSAPL1, CLEC10A, OR2G2, DISP1, VWA5B1, OR6A2, NUP210, UPK1A, BTN1A1, GALNS, UGT2A2, EGF, PCSK6, ST6GAL2, CPA6, FOXRED2, OR6C74, STAB2, GPR151, SERPINI2, PXDNL, LAMA4, SLC4A11, STAB1, SLC17A1, P2RX2, DSC2, ADRA1A, ABCC2*	**$3.4x10^{-3}$**

Continue on next page

Continued

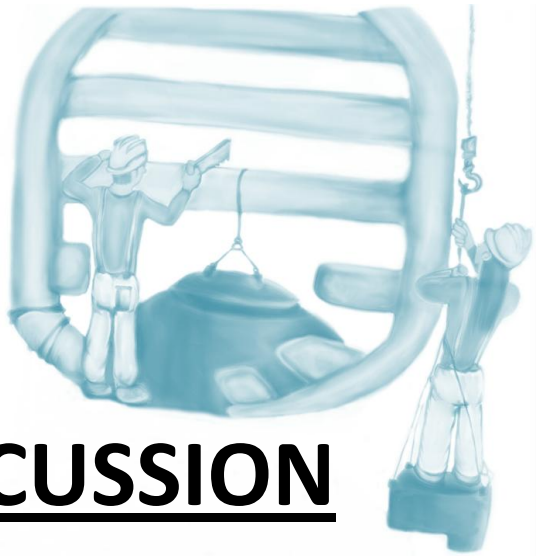
Category	Term	Count	*P*	Genes	P_{FDR}
UP_SEQ_FEATURE	Topological domain: cytoplasmic	59	$1.5x10^{-4}$	*PVR, QRFPR, SLC8A3, OR10A4, SLC44A2, GBGT1, TRPV3, CHPF2, RAET1E, CHST15, ANO7, ATP8B3, LAG3, CEACAM20, ATG9B, TRPM5, ATP4A, KCND1, PCDHB6, PCDHB4, BTNL2, PCDHGB6, SLC22A20, PKD2L1, TRPM1, LAX1, SUSD5, ERN2, OR6V1, GLRA4, GLP1R, HLA-DQB1, CALCR, FGFR4, SLC2A11, CLDN19, BTN2A1, OR51F2, CLEC10A, OR2G2, NUP210, UPK1A, OR6A2, BTN1A1, EGF, ST6GAL2, NOX5, OR6C74, GJB3, GJB4, STAB2, GPR151, OR10H3, SLC4A11, STAB1, P2RX2, ADRA1A,DSC2, ABCC2*	**0.028**
GOTERM_CC_FAT	Plasma membrane	107	$5.3x10^{-5}$	*LMO7, OR8S1, GRIN3B, LPHN2, OR4D2, CHRNA9, CD44, WNK4, GRID1, CLCA1, OR10S1, MYH1, PTPRN2, CDHR2, HLA-C, CTNNA3, LILRB4, GPR50, OR8B4, KBTBD10, DST, SNTG2, HDLBP, FGFR4, PANX1, AMTN, OR2T1, KEL, ENPP3, MAP4K2, PAQR7, ITGAM, KCNS3, CRB2, PLCH2, P2RY2, TRO, BAI2, B3GNT3, EGF, SLC28A3, ADAM28, EPB41, LPP, KCNB1, ITGA3, OR5AC2, GPR35, SLC4A11, EPS8, P2RX3, NOTCH4, TSC2, MAP7, OR51A7, PLA2G4E, CHRNG, TM7SF4, CLSTN2, CLSTN3, KCNJ12, CDSN, FCRL3, NMUR2, SDPR, OR6P1, KCNG4, RHOF, USH2A, OR2AE1, TRPM2, OR10J1, NCR3, CD86, TNFRSF10C, CLECL1, SLC26A9, GRM6, PLA2G3, NKD2, LRRC32, TAP2, COL6A2, COL6A1, CCBP2, PLA2R1, SLC39A4, OR8G2, PPP1R9B, COG4, OR51M1, CDH13, PROM2, ERBB2IP, NRAP, GPR110, SLC17A2, GPR113, MEP1B, ANXA13, GPR111, SYNM, TAPBPL, ABCC8, GFRA2, ABCC6, GPR116*	**$1.5x10^{-2}$**

Continue on next page

Annotation cluster 2	**Enrichment score: 2.15**				
Category	**Term**	**Count**	***P***	**Genes**	**P_{FDR}**
UP_SEQ_FEATURE	Topological domain: cytoplasmic	59	$1.4x10^{-4}$	*PVR, QRFPR, SLC8A3, OR10A4, SLC44A2, GBGT1, TRPV3, CHPF2, RAET1E, CHST15, ANO7, ATP8B3, LAG3, CEACAM20, ATG9B, TRPM5, ATP4A, KCND1, PCDHB6, PCDHB4, BTNL2, PCDHGB6, SLC22A20, PKD2L1, TRPM1, LAX1, SUSD5, ERN2, OR6V1, GLRA4, GLP1R, HLA-DQB1, CALCR, FGFR4, SLC2A11, CLDN19, BTN2A1, OR51F2, CLEC10A, OR2G2, NUP210, UPK1A, OR6A2, BTN1A1, EGF, ST6GAL2, NOX5, OR6C74, GJB3, GJB4, STAB2, GPR151, OR10H3, SLC4A11, STAB1, P2RX2, ADRA1A, DSC2, ABCC2*	**0.028**
UP_SEQ_FEATURE	Topological domain: extracellular	50	$8.8x10^{-4}$	*PVR, QRFPR, SLC8A3, SLC44A2, OR10A4, TRPV3, RAET1E, ANO7, ATP8B3, LAG3, CEACAM20, TRPM5, PCDHB6, PCDHB4, BTNL2, PCDHGB6, SLC22A20, PKD2L1, TRPM1, LAX1, SUSD5, OR6V1, GLRA4, GLP1R, CALCR, HLA-DQB1, FGFR4, SLC2A11, CLDN19, BTN2A1, OR51F2, CLEC10A, OR2G2, UPK1A, OR6A2, BTN1A1, EGF, NOX5, GJB3, OR6C74, GJB4, STAB2, GPR151, OR10H3, SLC4A11, STAB1, P2RX2, ADRA1A, DSC2, ABCC2*	**0.029**

Gene-set enrichment analysis of gene-based P-values from SKAT-O was performed using the functional annotation clustering analysis module of the bioinformatic tool DAVID. Each annotation term group is assigned an enrichment score to rank overall importance. Only annotation clusters with ES≥1.3 (indicating biological significance) and significant P-values after FDR correction are shown. The significance of gene-term enrichment was assessed with a modified Fisher´s exact test and P-values are corrected using Benjamini-Hochberg's by false discovery rate (FDR-BH) procedure**.**

Continued



DISCUSSION

Progress made in oncology treatment over the last half of the century has led to an increase in cure rates and overall survival. Although more patients are surviving cancer than ever before and even with current therapies, some cancer patients, such as those with metastatic and/or recurrent/refractory Ewing sarcoma[58], have a dismal prognosis; and/or suffer ADRs which compromise treatment efficacy and patient´s quality of life (such as CiHFS[84,85]) and many are left with irreversible, life-threatening and long-lasting toxicities (such as chronic AIC[104]). The prediction of drug efficacy and drug side effects to allow therapy individualization is very important in cancer chemotherapy, given that antineoplastic agents are, in general, nonspecific with a narrow therapeutic index[5]; and cancer incidence is expected to rise over the next several decades[194].

Large interpatient variability observed in both, the efficacy and toxicities associated with chemotherapeutics drugs, although is multifactorial can be explained in part by the host genetic background. In this thesis we particularly focused on the identification of germline genetic variants that predict treatment outcome in children diagnosed with Ewing sarcoma, an aggressive pediatric tumor with a huge unmet need to improve the outcome in patients with metastatic disease and also in the recurrent setting[58] (**Study I**). In addition, we focused on the identification of genetic variants associated with increased risk to develop CiHFS (a common skin adverse event and the most frequent cause for capecitabine dose reduction or therapy discontinuation[84]) (**Study II**) and chronic AIC (a serious adverse drug reaction limiting anthracycline use and causing substantial morbidity and mortality[114]) (**Study III** and **Study IV**).

1. Identification of genetic variants in pharmacokinetic genes associated with Ewing Sarcoma treatment outcome (Study I)

Over the last four decades, the outcome of patients with localized Ewing sarcoma has improved considerably; however, drug-resistant disease at diagnosis or at relapse remains a major cause of mortality among pediatric patients diagnosed with Ewing sarcoma[56,140]. Although candidate gene approaches have been broadly used in oncology and numerous studies of somatic prognostic biomarkers have been performed in Ewing sarcoma; there is a lack of studies exploring how germline genetic variation in genes that play an important role in the pharmacokinetics of the commonly used cytotoxics in Ewing sarcoma therapy would affect both, treatment efficacy and patients´ overall survival. Owing to these, we performed the first study considering an integrated pathway-based approach and assessed associations with treatment outcome in a discovery cohort of 106 Spanish children diagnosed with Ewing sarcoma with

replication in a large independent cohort of 389 pediatric Ewing sarcoma patients from across Europe.

We identified variants significantly associated with tumor response in the discovery cohort, but we were not able to replicate these associations in the European cohort. This lack of association with tumor response could be explained by treatment heterogeneity in the neoadjuvant setting within each cohort and between patient series. However, including neoadjuvant therapy as an additional covariate in statistical analyses between genotypes and tumor response made no substantial difference to the results obtained. Lack of association could be also explained by considering together patients with localized and metastatic disease. We decided to consider metastatic patients and non-metastatic patients together, but including metastasis as covariate, to increase both the sample size and the statistical power; but still sample size might be underpowered to detect associations with tumor response.

In addition, we tested associations between SNV genotypes and overall survival. We identified and replicated the associations for three common variants, rs7190447, rs4148737 and rs11188147, located in *ABCC6*, *ABCB1* and *CYP2C8* genes, respectively. Both *ABCC6* and *ABCB1* genes are members of the ATP-binding cassette (ABC) transporter superfamily. These transporters translocate a variety of substrates, including chemotherapeutic agents through biomembranes against a concentration gradient, with ATP hydrolysis providing the driving force. The major physiological role of ABC transporters is to protect normal cells and tissues against exogenous toxins, thereby also affecting to the uptake and distribution of anticancer drugs in the human body. Importantly, overexpression of some ABC members, such as *ABCB1*, *ABCC1* or *ABCG2* contribute to chemoresistance by active extrusion of cytotoxic agents from the tumor cell[195,196]. Little is known about the physiological role of *ABCC6,* including its natural substrates or its role in drug excretion[197]. *ABCC6* is mainly expressed in the basolateral membrane of hepatocytes and in proximal kidney tubules[198]. Basolateral efflux transporters mediate the removal of endogenous and xenobiotics compounds from hepatocytes into blood. The role of ABC transporters located at the basolateral membrane of hepatocytes in detoxification of drugs it has been suggested to be minor compared to proteins located at the canalicular membrane, but is not well understood. It has been found that the expression of basolateral ABC proteins, such as ABCC1 and ABCC3 is increased when hepatic bile excretion is blocked (cholestasis), favoring renal elimination[198,199]. In contrast to *ABCC1* and *ABCC3*, *ABCC6* is significantly expressed in normal liver and its expression is not induced in ABCC2-deficient rats or during

cholestasis. Given this constitutive expression, a "housekeeping" function has been proposed for *ABCC6* in hepatocytes[200], but it has to be further explored. On the other hand, the role of *ABCC6* in renal drug handling remains to be defined[201]. Regarding the role of this gene in multidrug resistance, *ABCC6* can mediate the transport of several chemotherapeutic agents including etoposide, doxorubicin and actinomycin-D (drugs administered to Ewing sarcoma patients) and cisplatin, although the levels of resistance are significantly lower to those observed for *ABCC1* and *ABCC2*[202,203]. In addition, RNA expression profiling revealed that *ABCC6* is expressed in tumors of patients with localized Ewing sarcoma (but at low levels compared to liver and kidney)[204], so a role for *ABCC6* in multidrug resistance of Ewing sarcoma cells could not be ruled out. To date, it remains to be determined whether intronic polymorphisms in *ABCC6* have an impact on gene expression and, hence their effect on drug disposition[205,206]. Recent studies have demonstrated that variants in non-transcribed regions can influence gene expression through regulatory mechanisms[207]. ENCODE and HaploReg data suggest that the genomic region containing rs7192303 (in perfect LD with our replicated variant, rs7190447) might be regulatory. Based on these data, the SNV rs7192303 can potentially affect CTCF and cohesin binding to DNA. The mammalian genome is organized by large regions [topologically associating domains (TADs)] defined by high levels of chromatin interaction within the region and little or no interaction with neighbouring regions[208,209]. Approximately 70% of chromatin looping events, which facilitate gene regulation by distant regulatory elements, occur within TADs. The typical length scale of chromatin looping is 10–200 kb, whereas that of TADs can be as large as 2000 kb[210]. CTCF is a highly conserved protein that has been classically considered an insulator[211]; whereas cohesin is a ring-shaped complex with a crucial role in establishing sister chromatid cohesion[212]. In addition to these well-established functions, CTCF and cohesin colocalize and cooperate to control gene expression and genomic organization. CTCF is involved in promoting long-range interactions and DNA looping and TAD boundaries are typically defined by CTCF binding sites. A large number of cohesin-binding sites colocalize with CTCF-binding sites and it has been shown that cohesin is required for the stabilization of chromatin loops at certain loci. Both, cohesin and CTCF contribute to long-range chromatin contacts within the TADs[211,213,214]. CTCF-mediated chromatin interactions determined by chromatin interaction analysis with paired-end tag (ChIA-PET) in ENCODE have been analyzed to date in five different human cancer cell lines [(K562 (chronic myeloid leukemia), HCT116 (colorectal cancer), HeLa-S3 (cervical cancer), MCF-7 (breast cancer), and NB4 (promyelocytic)], but none of them are hepatic or Ewing sarcoma cell lines. However, ChIA-PET analysis in MCF-7 cells revealed that the rs7192303 overlapping region has the potential to form a CTCF-mediated chromatin loop with an upstream intronic region of *ABCC6*

gene and with an intronic region of the upstream gene *ABCC1.* Interestingly, rs7192303 is an eQTL in the liver. Although ChIP-seq for CTCF and cohesin or 4C-seq experiments in hepatocytes and Ewing sarcoma cells should be performed in order to confirm the specific interactions affected by rs7192303; we hypothesize that differences in *ABCC6* expression caused by changes in CTCF and/or cohesin binding due to rs7192303 could affect the efflux of *ABCC6* target-drugs used in Ewing sarcoma standard treatment, thus affecting systemic bioavailability of anticancer agents and intracellular drug levels in tumor cells; which ultimately determines treatment response and patient survival.

ABCB1 (P-glycoprotein) was the first ABC transporter identified and has become the most studied gene in the field of multidrug resistance. *ABCB1* is expressed in the intestine, liver, kidney, the blood-brain and placental barriers, with apical membrane localization. The expression pattern and the broad substrate specificity, make *ABCB1* a major player in drug absorption, disposition and elimination[199,215]. Of the drugs administered to Ewing sarcoma patients, etoposide, vincristine, doxorubicin and actinomycin-D are transported by this pump[216]. *ABCB1* mRNA and/or protein expression has been frequently detected in tumor samples and reported to be associated with clinical outcome in several reports[216,217]. Along with *ABCC1, ABCB1* is the only ABC gene that has been investigated in some detail in Ewing sarcoma; however, the findings of published studies are contradictory with two suggesting that ABCB1 protein expression in Ewing sarcoma tumors is not predictive of prognosis[218,219], and a third significantly linking protein expression in tumors to poor response to therapy[220]. Genetic polymorphisms in *ABCB1* have been reported to change mRNA/protein expression and function, however little attention has been given to intronic and non-coding variants in this gene, and their possible link to cancer[221]. Consistent with a previous study performed by my own group in which a significant association for the minor G allele of rs4148737 with poorer overall survival in pediatric osteosarcoma patients was reported[141], in the current thesis we observed that the GG genotype was associated with higher risk of death, suggesting that rs4148737 may be important as a prognostic marker after treatment in the two most common pediatric bone tumors (osteosarcoma and Ewing sarcoma). In addition, rs4148737 was predicted to affect RUNX3 binding to DNA. It has recently been shown that Ewing sarcoma tumors express RUNX3 and that RUNX3 binds to EWS/FLI. In addition, suppression of RUNX3 causes a reduced growth of Ewing sarcoma cell lines and disrupts the expression of EWS/FLI-regulated genes, indicating an oncogenic role for RUNX3 in Ewing sarcoma tumors[222].

CYP2C8 is a member of the human CYP2C enzyme family, which collectively are significant contributors to drug disposition and are responsible for the metabolism of about 20% of clinically available drugs. It is highly expressed in liver, but is also found in extrahepatic sites such as the kidney, brain, uterus, mammary gland, ovary, heart, adrenal gland and duodenum. *CYP2C8* plays a major role in the oxidative metabolism of many drugs (e.g., thiazolidinediones, meglitinides, nonsteroidal anti-inflammatory drugs), but also chemotherapeutics, including two of those used in Ewing sarcoma treatment (cyclophosphamide and ifosfamide)[223–226]. Although it has been shown that there is great interindividual variation in the metabolism of *CYP2C8*-specific substrates and in *CYP2C8* expression[223,227], nothing has previously been reported about the impact of *CYP2C8* polymorphisms and their implications for clinical outcome in patients treated with cyclophosphamide and ifosfamide.

While one might expect the three replicated genetic variants also to be associated with tumor response, we didn´t observe this association. The evaluation of tumor response after administration of Ewing sarcoma neoadjuvant therapy, when chemotherapy treatment had not been completed, could explain the lack of observed association with this clinical feature, particularly bearing in mind that ABCB1 and ABCC6 not only transport neoadjuvant drugs but also adjuvant chemotherapeutics, and CYP2C8 is in part responsible for the oxidative metabolism of neoadjuvant and adjuvant Ewing sarcoma agents. Low statistical power because of low sample size of the study could be also a reason for the failure to identify associations.

In conclusion, we have carried out a two-stage study using a multi-drug transport/metabolism pathway approach, and identified for the first time germline genetic variants in the *ABCC6*, *ABCB1* and *CYP2C8* genes significantly associated with overall survival in Ewing sarcoma patients. The results of our study, which have been replicated in a large cohort of patients, emphasize the clinical relevance of these genes as prognostic marker genes, although experimental verification of putative regulatory function for the identified variants will be required.

2. Identification of genetic variants predictive of susceptibility to capecitabine-induced hand-foot syndrome (CiHFS) (Study II)

CiHFS is a common dose-limiting toxicity of capecitabine, occurring in more than 30% of patients. If not promptly managed, CiHFS can progress to an extremely painful and disabling

condition (grade 3 toxicity), causing impairment of function and significant discomfort, leading to treatment withdrawal, dose reduction and worsened quality of life of these patients (~17% of capecitabine-treated patients)[80,84,85].

The most comprehensive analysis performed to date of capecitabine toxicity pharmacogenetics is the QUASAR2 (Quick and Simple and Reliable trial) study[97]. They evaluated associations between previously reported variants (36 variants), all of them in genes involved in the biochemical pathway of capecitabine activation, action or degradation; with common capecitabine toxicities, including CiHFS, in 927 colorectal cancer patients treated with standard capecitabine regimen following surgery resection of stage II/III tumors. They also performed a meta-analysis of QUASAR2 and 16 published studies (N=4,855 patients) (included one study focused on candidate genes performed by my own group[93]) to examine candidate polymorphisms in capecitabine and 5-FU monotherapy but also in combination therapy protocols. They found only four variants: *DPYD* 2846TA, *DPYD**2A, *TYMS* 5´ VNTR 2R/3R and *TYMS* 3´ UTR 6bp ins-del, significantly associated with global high grade capecitabine toxicity, especially those variants in the DPYD enzyme (combined OR: 5.51). These variants remain significantly associated when CiHFS was the only toxicity evaluated, but none of them was associated with global or any specific toxicity when combination regimens were administered. Interestingly, when they assessed the prediction capacity of global toxicity of *DPYD* and *TYMS* variants, they found that these variants had only a modest power to predict capecitabine toxicity. These results underscore that the interindividual variability in CiHFS susceptibility remains largely unexplained and the urgent need to identify valuable predictive markers to accurately stratify patients at high risk for CiHFS, but also to elucidate the precise molecular mechanisms underlying this common ADR. By combining GWAS analysis, fine-scale mapping, and functional analyses using human skin cells and tissues, we provide for the first time compelling evidence that *CDH4* regulatory variants are involved in the development of CiHFS. More important, we found that levels of R-cadherin and involucrin are lower in the skin of patients who go on to develop severe CiHFS compared to patients who don't, prior to capecitabine treatment; uncovering a novel mechanism of CiHFS susceptibility via the perturbation of the skin barrier function.

Of note is in this study we considered exclusively patients showing extreme phenotypes; those patients suffering the most severe form of CiHFS toxicity (grade 3) and those not experiencing any toxicity during capecitabine treatment (grade 0). Given that CiHFS toxicity is

probably a polygenic trait and is measured using a scale-classification system, selecting most-sensitive and most-resistant patients it deemed essential to evaluate only highly informative patients having an unequivocal phenotype; but also the use of this extreme phenotype sampling enriches the presence of causal variants in the patients, increasing the probability of discovering associations. On the other hand, patients included in the study were treated with capecitabine monotherapy, avoiding overlapping toxicities and interactions between drugs.

The vast majority of disease-associated SNVs identified by using commercial probe-based genotyping array platforms are located in non-coding regions of the genome, equally proportioned between the intergenic and intronic regions[36,228,229] and many are expected to be eQTLs[230]. Thus, it is likely that the underlying mechanism linking them to the disease is through gene regulation. Schaub et al.[231] studied 4,724 SNVs from the GWAS catalog[36] associated with 470 different phenotypes using ENCODE data, and found that 36% of the variants are in DNase-hypersensitive regions (regions that correspond to areas of open, accessible chromatin that contain binding motifs for transcription factors) and 20% are in a ChIP-seq cluster in at least one cell line. Ernst et al.[232] generated a genome-wide map of nine chromatin marks across nine cell types to systematically characterize regulatory elements. They found a 2-fold enrichment for predicted strong enhancers among the associated SNVs from 426 GWAS and several of these variants created or disrupted transcription factor motifs in the identified enhancers. However, understanding how allele-specific genetic variation affects gene expression requires a complete and cell type specific picture of both the spatial organization and the functional features of each locus. Tang et al.[233] used the ChIA-PET technique to comprehensively map the chromatin organization and specific interactions mediated by CTCF and also by RNA polymerase II. They found that SNVs located at chromatin contact boundary regions and within the core CTCF binding motif, can abrogate CTCF binding, looping, and chromatin topology. They also found that several of disease-associated SNVs resides in CTCT motifs, dramatically altering chromatin organization, sustaining the possibility that changes in chromatin topology driven by SNVs could be the primary molecular event underlying disease susceptibility. Interestingly, only 25% of SNVs found in 51 different lymphoblastoid cell lines genotyped as part of the 1000 Genomes project were located exactly in the core CTCF binding motif, but the vast majority were in nearby regions (within 1 kb of the motif)[234]. Although is not yet known the functional relevance for chromatin topology of SNVs that are nearby but not exactly map with CTCT motifs; several works suggest that flanking sequences outside the core binding motif can profoundly affect transcription factor binding to DNA (through affecting DNA shape properties such as minor groove width, roll; or

composition of poly(dA:dT) tracts)[235,236]. The four highly correlated variants associated with susceptibility to CiHFS identified in our study are nearby to CTCF or cohesin binding sites, but they did not overlap exactly with the core motifs; however, our 4C-seq experiments revealed a physical interaction between the *CDH4* promoter and the risk alleles containing region in the presence of the risk alleles. This interaction was clearly decreased in a different keratinocyte cell line homozygous for the reference haplotype. Although this result is not sufficient to conclude that the presence of the risk alleles is responsible for the observed changes in the contact profile, it certainly points in this direction. One interesting possibility would be that the presence of these risk variants creates new binding sites of some of the aforementioned factors, generating a contact that negatively regulates *CDH4* expression; although ChIP-seq experiments for CTCF and cohesin should be performed to explore this hypothesis. Importantly, we found that the haplotype carrying the risk alleles produced a significant decrease in *CDH4* mRNA expression and correlated with reduced protein levels.

CDH4 encodes R-cadherin. Cadherins constitute a large family of cell surface proteins, many of which participate in Ca^{2+}-dependent cell-cell adhesion through the establishment of adherens junctions and desmosomes; playing a fundamental role in the preservation of proper tissue architecture and function[237,238]. The epidermis is composed of four functionally different layers of keratinocytes at different stages of differentiation, which undergo programmed differentiation to allow for constant renewal of the skin. Epidermal differentiation begins with the migration of keratinocytes from the basal layers, and ends with the formation of the cornified layer. Cell proliferation, differentiation and death occur sequentially, and each process is characterized by the expression of specific proteins, and specific diseases can result from abnormalities in these proteins[191]. Epidermal keratinocytes express two classic cadherins: E- and P-cadherin, with different epidermal expression patterns. While E-cadherin is found in all epidermal layers, P-cadherin is restricted to the basal layer, the innermost layer of the epidermis with keratinocytes in a proliferative stage[239,240]. On the other hand, mutations in cadherins are known to cause inherited skin disorders, such as the ectodermal dysplasia, ectrodactyly, and macular dystrophy syndrome or striate palmoplantar keratoderma, which is characterized by a thickening of the skin on the palms of the hands and the soles of the feet[184]. R-cadherin is highly expressed in brain and plays an essential role during brain segmentation and in neuronal outgrowth; and is important for kidney and striated muscle development[241,242]; however whether or not R-cadherin plays a role in skin physiology or pathology has not yet been investigated. We found that R-cadherin distribution in human epidermis differed from that for E and P-cadherin, with R-cadherin mainly localized in the suprabasal granular layer of the human

epidermis, where keratinocytes are in a differentiated stage. That R-cadherin is highly expressed by differentiated keratinocytes was confirmed by *in vitro* assays. Interestingly, we observed that decreased R-cadherin expression leads to a decrease of involucrin, at both mRNA and protein levels, but not affecting others differentiation markers. Involucrin is a structural protein of the cornified envelope, the outermost layer of the epidermis composed of terminally differentiated keratinocytes (corneocytes). The cornified envelope consists of keratins enclosed within tightly knit proteins, which are crosslinked by transglutaminases and surrounded by a lipid envelope; and is essential for the mechanical integrity and water impermeability of the skin[191,238]. Several lines of evidence suggest that involucrin is an early component in the assembly of the cornified envelope[243,244], acting as a scaffold, to which other proteins such as cystatin α, elafin, small proline-rich proteins, and loricrin subsequently become crosslinked to complete the cornified envelope assembly. In addition, involucrin might be the preferred protein substrate for the covalent attachment of ceramides to form the exterior surface of the cornified envelope[191,238]. Surprisingly, mice lacking individual components of the cornified envelope, such as loricrin[245], involucrin[246], periplakin[247] and envoplakin[248], developed normally and possessed apparently normal epidermis and hair follicles and no obvious cornifed envelope abnormalities; while the combined loss of involucrin, envoplakin, and periplakin[249] impairs the epidermal barrier; suggesting a compensatory redundancy between these proteins. Although involucrin deficiency may not directly alter skin function, consistent with the lack of symptoms we observed in the skin of the patients prior capecitabine treatment; mutations in the granular layer and cornified envelope components are thought to contribute to genetic susceptibility to chronic barrier defects, as in atopic dermatitis[250,251]; and altered expression of involucrin has been found in skin diseases characterized by disturbance of cornification, such as psoriasis[252] and lamellar ichthyosis[253]. We hypothesize that reductions in involucrin may increase the cytotoxic effect of capecitabine leading to a breakdown of the skin epidermal barrier. This effect could be particularly dramatic in the palms of the hands and soles of the feet, due to their higher epidermal cell division rate, greater expression of capecitabine-metabolizing enzymes (including TYMP and DPYD), leading to the accumulation of capecitabine metabolites and catabolites; increased vascularization, pressure and temperature; as well as increased local drug elimination by the eccrine glands of these areas[84,87].

Taken together, our results reveal a novel implication of R-cadherin in the regulation of keratinocyte differentiation through involucrin expression, although further studies are needed to explore how R-cadherin goes beyond this structural function to regulate gene expression.

Importantly, R-cadherin and involucrin levels present in the skin of the patients sampled prior to capecitabine treatment inversely correlate with CiHFS upon treatment, which could be useful for patients risk stratification.

In summary, using a multifaceted approach, we have unearthed a novel risk locus strongly associated with CiHFS susceptibility. Our results open new and unprecedented avenues for future research, to deepen our knowledge of the pathogenic mechanisms underlying CiHFS, and provide novel insight into the clinical application of the identified risk variants along with the R-cadherin and involucrin expression as predictive biomarkers of CiHFS as a further step towards personalized cancer treatment.

3. Role of low-frequency variants in susceptibility to chronic anthracycline–induced cardiotoxicity (AIC) (Study III and Study IV)

Life-threatening or permanently disabling complications of therapy can occur despite the administration of medications at the recommended dose. Long-term cardiotoxicity is a well known complication of anthracycline treatment; however, clinicians remain unable to accurately stratify patients into high or low-risk groups for AIC and genetic factors influencing AIC susceptibility are still largely unexplained. In order to explore the contribution of coding variants, especially low-frequency and rare variants, to susceptibility to AIC we performed an exome array analysis considering variants on the Illumina HumanExome BeadChip array. Considering that children are particularly vulnerable to the cardiotoxic effect of anthracyclines, even more than adults and/or at lower doses, in Study III we focused on the identification of susceptibility variants in anthracycline-treated pediatric patients diagnosed with Ewing sarcoma, osteosarcoma or leukemia; while in Study IV, we focused on adult patients diagnosed with breast cancer, but also in pediatric patients.

Regarding the demographic and clinical characteristics of the 93 anthracycline-treated pediatric cancer patients (Study III), of note is that controls were significantly younger than cases and gender distribution in our cohort showed more female patients in controls than in cases patients, although it is well known that younger patients and girls are particularly vulnerable to AIC. Remarkably, osteosarcoma and Ewing Sarcoma were primary diagnosis significantly more common in cases than in controls. High rates of osteosarcoma and Ewing Sarcoma among cases, may be explained because patients diagnosed with this type of pediatric bone tumors received higher cumulative anthracycline doses, than those diagnosed with leukemia [median cumulative anthracycline dose (mg/m^2)= 446.2 (osteosarcoma); 361.2 (Ewing sarcoma) and 132 (leukemia)].

Although we decided not to include the tumor type as a covariate in the corresponding statistical analyses given the greater anthracycline doses of solid tumors in our series and even we know about the multicollinearity issues; inclusion of primary diagnosis as an additional covariate in multivariable models made no substantial difference to the results obtained. Regarding the demographic and clinical characteristics of the breast cancer patients in Study IV, more patients in the Belgium cohort experienced chronic AIC compared to the Spanish breast cancer cohort, although Belgium patients were treated with epirrubicin, which has improved cardiac tolerability compared to doxorubicin[109].

Using standard single-variant association tests, none of the 246,060 variants analyzed was found to be statistically significantly associated with chronic AIC after correction for multiple comparisons, neither in pediatric patients (Study III) nor in breast cancer patients (Study IV). Because the number of rare variants is much larger and rare variants are less correlated with each other than common variants, more stringent significance thresholds that those typically employed in GWAS are required, reducing the statistical power (especially in cohorts with small sample size and few patient cases, as our patient series)[53,254,255]. Instead of testing each variant individually, we focused on gene-based tests that have greater statistical power to detect associations with rare variation and can evaluate the cumulative effect of multiple genetic variants (including both, common and rare) within a gene[255]. As it was mentioned in the introduction, some genes/regions may have a high proportion of causal variants and influence the phenotype in the same direction while others may have few causal variants or the causal variants may have different directions of association. Therefore, the use of methods optimal for both scenarios, such as the combined gene-based test SKAT-O applied in Study III and IV, is desirable[55,166].

Using this kind of approach, we have identified *GPR35* as the gene most significantly associated with chronic AIC in children (Study III). *GPR35* belongs to the G-protein-coupled receptor family, which are membrane proteins mediating a wide range of physiological processes[256]. Although the exact functions of *GPR35* are not known, several lines of evidence strongly suggest potential roles for this receptor in cardiac physiology and pathology. Sun et al.[257] were the first to report a cardiovascular role for *GPR35*, with a common non-synonymous SNV (rs3749172) significantly associated with the burden of coronary artery calcification, a subclinical measure of atherosclerosis and coronary artery disease. Importantly, this variant, rs3749172, is the only genetic variant in *GPR35* reported to be associated with cardiovascular disease to date, and no further studies have explored the contribution of genetic variants in

GPR35 to cardiovascular diseases. rs3749172 is covered by the exome array we used, however, this variant was not significantly associated with chronic AIC in children (OR=1.82, *P*=0.148, P_{FDR}=0.96). *GPR35* was later found to be up-regulated in failing myocardium of 12 patients with severe chronic heart failure by performing gene expression microarray analysis, and in the same study *GPR35* knock-out mice showed higher systolic blood[258]. More recently, *GPR35* has been characterized as a novel hypoxia-sensitive gene, with hypoxic conditions (mediated by hypoxia-inducible factor 1 activation, *HIF-1*) causing increased expression at both mRNA and cell surface protein levels in neonatal mice cardiomyocytes. In a mouse surgical model of acute myocardial infarction and hypertrophy *GPR35* behaved as a typical hypoxia response gene: its expression increased in early phases of cardiac infarction, but unaltered at later phases; suggesting that *GPR35* is an early marker of progressive cardiac failure[259]. In vitro functional assays in neonatal rat cardiomyocytes demonstrated that *GPR35* overexpression reduced cell viability and caused hypertrophy[204], whereas in neonatal mouse cardiomyocytes promoted morphological changes, including membrane ruffling and retraction fibre formation[259]. Remarkably, several studies have linked *GPR35* to inflammatory regulation[260,260] and there is ample evidence to support the hypothesis that inflammation, as well as hypoxia, play a significant role in the pathogenesis and development of chronic cardiac complications, including cardiomyopathy[261,262]. These findings provide a possible explanation for the involvement of this receptor in cardiovascular disease. On the other hand, the cellular and biological effects of *GPR35* on cardiovascular pathophysiology could be largely mediated by downstream signaling pathways such as $G\alpha_{13}$, $G\alpha_{i/o}$ and RhoA, following receptor activation[263–266]. Interestingly, putative endogenous ligands of *GPR35* have also been linked to cardiovascular disease: lysophosphatidic acid has been associated with regulation of blood pressure and atherosclerosis[267]; and levels of kynurenic acid and reverse thyroid hormone T3 were found to increase (and T3 to decrease) in spontaneously hypertension rats[268] and in patients with chronic advanced heart failure[269] or acute myocardial infarction[270], respectively. While the SKAT-O does not provide any parameter estimates, like OR in logistic regression, we assessed the individual contribution of variants within *GPR35*. We found that rs12468485, a missense low-frequency variant (p.Thr253Met), strongly associated with the most severe cardiac manifestations (LV dysfunction, mostly symptomatic, evidenced after treatment with anthracycline doses well below the average for cases), was most influential. Although in silico prediction algorithms yielded discrepant results, rs12468485 was predicted to a have a potentially regulatory role in splicing. Due to the incomplete coverage of coding variants on the exome array, we sequenced the coding region of *GPR35* in our series of 93 anthracycline-treated

pediatric cancer patients but we were unable to identify additional independent association signals at the gene.

On the other hand, gene-based analysis in anthracycline-treated breast cancer patients revealed that variant alleles of low-frequency missense variants in *ETFB* and *WISP1* genes significantly increased chronic AIC risk (Study IV). ETFB is the β subunit of the heterodimer electron transfer flavoprotein (ETF) protein located in the inner mitochondrial membrane. ETF acts as an electron acceptor of energy production from amino acid and fatty acids that transfers electrons to the main respiratory chain via the ETF-ubiquinone oxidoreductase (ETF-QO)[271]. Anthracycline therapy is known to inhibit long chain fatty acid oxidation and transport across mitochondrial membrane[272] and decrease ATP production[273]; and oxidative stress has been proposed as a major contributor to anthracycline mediated myocardial injury[109,274]. Analysis of protein expression in doxorubicin-treated adult rat cardiomyocytes revealed differential downregulation of *ETFB*[275]. In addition, proteomic analyses of cardiac proteins from mice treated with doxorubicin showed elevated oxidative modifications of cardiac proteins, including ETF-QO, and these oxidative modifications altered their enzymatic activity[276], thus compromising ATP production in cardiac mitochondria. Taken together, these findings indicate that *ETF* is an important target in anthracycline-mediated oxidative stress. Remarkably, we found that rs79338777 variant allele carriers were scarce among control patients in the three independent cohorts we assessed, and the variant allele was significantly associated with higher risk of chronic ACT when the two breast cohorts and the pediatric series were analyzed together. This low-frequency missense variant was predicted to have a pathogenic effect by three of the in silico prediction algorithms used, including the consensus classifier.

WISP1 is a member of the CYR61/CTGF/Nov family of growth factors which although expressed in the heart at low basal levels; it mediates cardiac remodeling after myocardial infarction[277] and is upregulated in postinfarct myocardium[278]. More recently, Venkatesan et al[193]. reported that WISP1 inhibits doxorubicin-mediated cardiomyocyte death by blocking doxorubicin-induced p53 activation, p38 MAPK and JNK phosphorylation, Bax translocation to mitochondria, and cytochrome *c* release into cytoplasm. Collectively, these findings strongly indicate a pro-survival function for *WISP1,* antagonizing cardiomyocyte death in response to various adverse and stress conditions; including anthracycline administration. Although rs72731540 was predicted to be non-pathogenic, the majority of variant allele carriers suffered chronic AIC in all the three cohorts and the variant allele exhibit a significant greater risk of AIC.

rs79338777 in *ETFB* and rs72731540 in *WISP1* were associated with risk independently of the age of tumor onset, pointing out oxidative stress and the prosurvival function mediated by *WISP1* as shared molecular mechanisms between children and adults. On the contrary, the variant allele of rs12468485 in *GPR35* was almost exclusively present in cases in the Spanish pediatric cancer cohort, while it was found in controls in the Spanish breast cancer series. The opposite direction of association for rs12468485 between children and adults, point out the possibility that the molecular mechanism underlying AIC and mediated by *GPR35* could be exclusive or particularly relevant for cancer survivors of childhood cancers.

Both studies are limited in its ability to look at single-associations for very rare variants and detection of low-frequency variants with moderate effects due to a small sample size of the cohorts and the incomplete coverage of all coding variants at each locus provided by the exome array. However, the strengths of both studies included the well-characterization of the series of patients included in both studies, with extensive patient, tumor and therapy-related information; and notable long-term follow-up of all individuals.

In conclusion, we identified novel low-frequency coding variants associations and we extended the allelic spectrum of variation underlying chronic ACT susceptibility in children and in adults treated with anthracyclines. However, functional characterization of the observed associations and further replication in very large studies are required.

CONCLUSIONS

1. The studies carried out in this thesis have led to the identification of susceptibility genetic variants associated with treatment outcome in Ewing sarcoma patients and the development of adverse drug reactions in capecitabine-treated and anthracycline-treated cancer patients.

2. We have identified three germline variants in *ABCC6*, *ABCB1* and *CYP2C8* associated with overall survival in Ewing sarcoma patients. These associations were replicated in a large independent cohort, highlighting the importance of these pharmacokinetic genes as prognostic markers in Ewing sarcoma.

3. We have identified and replicated in an independent cohort a locus near the *CDH4* gene, which encodes R-cadherin, a protein mainly localized in the granular layer of the epidermis, strongly associated with the development of severe capecitabine-induced hand-foot syndrome.

4. We demonstrated that these risk variants prompted changes in *CDH4* gene expression, possibly through changes in chromatin topology; uncovering a novel mechanism underlying individual genetic susceptibility to capecitabine-induced hand-foot syndrome through impairment of keratinocyte differentiation and function and skin barrier disruption.

5. In the skin from breast cancer patients sampled prior to capecitabine administration, we have found that the levels of R-cadherin protein and involucrin, a protein of the cornified envelope essential for skin barrier function, were inversely correlated with the appearance of capecitabine-induced hand-foot syndrome.

6. We have identified *GPR35*, a gene with important roles in cardiac physiology and pathology, and in particular rs12468485, a missense low-frequency variant, as an independent risk factor for chronic anthracycline-induced cardiotoxicity in pediatric oncology patients treated with anthracyclines.

7. Using gene-based analyses, we have identified two novel genes: *ETFB* and *WISP1* associated with risk of developing chronic anthracycline-induced cardiotoxicity in breast cancer patients. Further analyses revealed that the low-frequency missense variants rs79338777 in *ETFB* and rs72731540 in *WISP1*, significantly increased risk of chronic anthracycline-induced cardiotoxicity, independently of whether patients were adults or children at diagnosis.

8. Our results demonstrate the utility of whole-exome genotyping arrays to detect low-frequency variants implicated in adverse drug events such as anthracycline-induced cardiotoxicity.

9. While replication of these findings in large independent series of patients is essential before these genomic biomarkers can be used in clinical practice for cancer patients, our results open up a promising new avenue for the individualization of treatment strategies.

CONCLUSIONES

1. Los estudios llevados a cabo en esta tesis doctoral nos han permitido identificar variantes genéticas asociadas con la respuesta al tratamiento y supervivencia en pacientes infantiles diagnosticados con sarcoma de Ewing y a la aparición de efectos adversos producidos por los quimioterapéuticos capecitabina (síndrome mano-pie) y antraciclinas (cardiotoxicidad crónica).

2. Hemos identificado 3 variantes germinales en los genes farmacocinéticos *ABCC6*, *ABCB1* y *CYP2C8* asociados con supervivencia global en pacientes oncológicos infantiles diagnosticados con sarcoma de Ewing. Dichas asociaciones fueron validadas en un cohorte independiente, destacando la relevancia de los genes implicados en la farmacocinética para el pronóstico de dichos pacientes.

3. Identificamos y replicamos un locus próximo al gen *CDH4,* el cual codifica para R-cadherina, una proteína altamente expresada en la capa granular de la epidermis, fuertemente asociado con la aparición de síndrome mano-pie en su forma más severa.

4. La presencia de las variantes de riesgo afecta a la expresión de *CDH4,* probablemente mediante la alteración de la topología de la cromatina. Dichos resultados, ponen de manifiesto un nuevo mecanismo molecular subyacente a la susceptibilidad para el desarrollo de síndrome mano-pie a través de la alteración de la diferenciación de los queratinocitos en la epidermis y la función de barrera de la misma.

5. Por otro lado, en pacientes diagnosticados con cáncer de mama, previo al tratamiento con el fármaco capecitabina, hallamos que los niveles en la epidermis de las proteínas R-cadherina e involucrina, una proteína esencial en la función de barrera del estrato córneo de la epidermis, se encontraban inversamente correlacionados con el desarrollo posterior de síndrome mano-pie en su manifestación más severa.

6. Además identificamos a *GPR35*, un gen con un papel esencial tanto en la fisiología como en la patología cardiaca; y en particular, la variante codificante poco frecuente rs12468485, como factores de riesgo independientes para cardiotoxicidad crónica producida por antraciclinas en pacientes oncológicos infantiles.

7. Por último, identificamos dos nuevos genes, *ETFB* y *WISP1,* asociados con un mayor riesgo de desarrollar cardiotoxicidad crónica en pacientes de mama tratados con antraciclinas. Dentro de estos genes, las variantes codificantes poco frecuentes rs79338777 en *ETFB* y rs72731540 en *WISP1* fueron identificadas como factores de riesgo independientemente de la edad del paciente en el diagnóstico.

8. Estos resultados demuestran la utilidad de los arrays de genotipado centrados en variantes exónicas para la identificación de variantes de riesgo poco frecuentes.

9. Para implementar en la clínica las variantes genéticas identificadas en esta tesis doctoral como biomarcadores genómicos, es por supuesto necesaria la validación de dichas variantes en cohortes de pacientes independientes con un gran tamaño muestral. Sin embargo, nuestros resultados constituyen un primer avance hacia la denominada medicina personalizada.

REFERENCES

1. Evans, W. E. & Johnson, J. A. Pharmacogenomics: the inherited basis for interindividual differences in drug response. *Annu. Rev. Genomics Hum. Genet.* **2,** 9–39 (2001).

2. Spear, B. B., Heath-Chiozzi, M. & Huff, J. Clinical application of pharmacogenetics. *Trends Mol. Med.* **7,** 201–204 (2001).

3. Pirmohamed, M. & Park, B. K. Genetic susceptibility to adverse drug reactions. *Trends Pharmacol. Sci.* **22,** 298–305 (2001).

4. Lazarou, J., Pomeranz, B. H. & Corey, P. N. Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. *JAMA* **279,** 1200–1205 (1998).

5. Relling, M. V. & Dervieux, T. Pharmacogenetics and cancer therapy. *Nat. Rev. Cancer* **1,** 99–108 (2001).

6. Evans, W. E. & Relling, M. V. Pharmacogenomics: translating functional genomics into rational therapeutics. *Science* **286,** 487–491 (1999).

7. Ventola, C. L. Role of pharmacogenomic biomarkers in predicting and improving drug response: part 1: the clinical significance of pharmacogenetic variants. *P T Peer-Rev. J. Formul. Manag.* **38,** 545–560 (2013).

8. Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin. Pharmacol. Ther.* **69,** 89–95 (2001).

9. Food and Drug Administration, HHS. International Conference on Harmonisation; Guidance on E15 Pharmacogenomics Definitions and Sample Coding; Availability. Notice. *Fed. Regist.* **73,** 19074–19076 (2008).

10. European Medicines Agency. ICH Topic E15 Definitions for genomic biomarkers, pharmacogenomics, pharmacogenetics, genomic data and sample coding categories. (2013).

11. Mehta, S. *et al.* Predictive and prognostic molecular markers for cancer medicine. *Ther. Adv. Med. Oncol.* **2,** 125–148 (2010).

12. Nalejska, E., Mączyńska, E. & Lewandowska, M. A. Prognostic and predictive biomarkers: tools in personalized oncology. *Mol. Diagn. Ther.* **18,** 273–284 (2014).

13. Paugh, S. W. *et al.* Cancer pharmacogenomics. *Clin. Pharmacol. Ther.* **90,** 461–466 (2011).

14. Filipski, K. K., Mechanic, L. E., Long, R. & Freedman, A. N. Pharmacogenomics in oncology care. *Front. Genet.* **5,** 73 (2014).

15. Rodríguez-Antona, C. & Taron, M. Pharmacogenomic biomarkers for personalized cancer treatment. *J. Intern. Med.* **277,** 201–217 (2015).

16. Kalia, M. Biomarkers for personalized oncology: recent advances and future challenges. *Metabolism.* **64,** S16-21 (2015).

17. Shastry, B. S. SNPs: impact on gene function and phenotype. *Methods Mol. Biol. Clifton NJ* **578,** 3–22 (2009).

18. https://www.cancer.gov/publications/dictionaries/cancer-terms.

19. Rosenblum, D. & Peer, D. Omics-based nanomedicine: the future of personalized oncology. *Cancer Lett.* **352,** 126–136 (2014).

20. Committee on the Review of Omics-Based Tests for Predicting Patient Outcomes in Clinical Trials, Board on Health Care Services, Board on Health Sciences Policy & Institute of Medicine. *Evolution of Translational Omics: Lessons Learned and the Path Forward*. (National Academies Press (US), 2012).

21. http://www.internationalgenome.org/.

22. https://cancergenome.nih.gov/.

23. http://icgc.org/.

24. Tai, H. L. *et al.* Thiopurine S-methyltransferase deficiency: two nucleotide transitions define the most prevalent mutant allele associated with loss of catalytic activity in Caucasians. *Am. J. Hum. Genet.* **58,** 694–702 (1996).

25. Deenen, M. J., Cats, A., Beijnen, J. H. & Schellens, J. H. M. Part 1: background, methodology, and clinical adoption of pharmacogenetics. *The Oncologist* **16,** 811–819 (2011).

26. Patnala, R., Clements, J. & Batra, J. Candidate gene association studies: a comprehensive guide to useful in silico tools. *BMC Genet.* **14,** 39 (2013).

27. Huang, R. S. & Ratain, M. J. Pharmacogenetics and pharmacogenomics of anticancer agents. *CA. Cancer J. Clin.* **59,** 42–55 (2009).

28. Cheok, M. H. & Evans, W. E. Acute lymphoblastic leukaemia: a model for the pharmacogenomics of cancer therapy. *Nat. Rev. Cancer* **6,** 117–129 (2006).

29. Dias-Santagata, D. *et al.* Rapid targeted mutational analysis of human tumours: a clinical platform to guide personalized cancer medicine. *EMBO Mol. Med.* **2,** 146–158 (2010).

30. Frampton, G. M. *et al.* Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nat. Biotechnol.* **31,** 1023–1031 (2013).

31. Shao, D. *et al.* A targeted next-generation sequencing method for identifying clinically relevant mutation profiles in lung adenocarcinoma. *Sci. Rep.* **6,** 22338 (2016).

32. Daly, A. K. Genome-wide association studies in pharmacogenomics. *Nat. Rev. Genet.* **11,** 241–246 (2010).

33. Daly, A. K. Using genome-wide association studies to identify genes important in serious adverse drug reactions. *Annu. Rev. Pharmacol. Toxicol.* **52,** 21–35 (2012).

34. Hong, E. P. & Park, J. W. Sample size and statistical power calculation in genetic association studies. *Genomics Inform.* **10,** 117–122 (2012).

35. Edwards, S. L., Beesley, J., French, J. D. & Dunning, A. M. Beyond GWASs: illuminating the dark road from association to function. *Am. J. Hum. Genet.* **93,** 779–797 (2013).

36. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42,** D1001-1006 (2014).

37. Mroziewicz, M. & Tyndale, R. F. Pharmacogenetics: a tool for identifying genetic factors in drug dependence and response to treatment. *Addict. Sci. Clin. Pract.* **5,** 17–29 (2010).

38. Pérez-Gracia, J. L. *et al.* Selection of extreme phenotypes: the role of clinical observation in translational research. *Clin. Transl. Oncol. Off. Publ. Fed. Span. Oncol. Soc. Natl. Cancer Inst. Mex.* **12,** 174–180 (2010).

39. Nebert, D. W. Extreme discordant phenotype methodology: an intuitive approach to clinical pharmacogenetics. *Eur. J. Pharmacol.* **410,** 107–120 (2000).

40. Perez-Gracia, J. L., Gloria Ruiz-Ilundain, M., Garcia-Ribas, I. & Maria Carrasco, E. The role of extreme phenotype selection studies in the identification of clinically relevant genotypes in cancer research. *Cancer* **95,** 1605–1610 (2002).

41. Treviño, L. R. *et al.* Germline genetic variation in an organic anion transporter polypeptide associated with methotrexate pharmacokinetics and clinical effects. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **27,** 5972–5978 (2009).

42. Maher, B. Personal genomes: The case of the missing heritability. *Nature* **456,** 18–21 (2008).

43. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461,** 747–753 (2009).

44. Pritchard, J. K. Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.* **69,** 124–137 (2001).

45. Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.* **95,** 5–23 (2014).

46. Ng, P. C. & Henikoff, S. Predicting the effects of amino acid substitutions on protein function. *Annu. Rev. Genomics Hum. Genet.* **7,** 61–80 (2006).

47. http://genome.sph.umich.edu/wiki/Exome_Chip_Design.

48. Huyghe, J. R. *et al.* Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. *Nat. Genet.* **45,** 197–201 (2013).

49. Igartua, C. *et al.* Ethnic-specific associations of rare and low-frequency DNA sequence variants with asthma. *Nat. Commun.* **6,** 5965 (2015).

50. Kozlitina, J. *et al.* Exome-wide association study identifies a TM6SF2 variant that confers susceptibility to nonalcoholic fatty liver disease. *Nat. Genet.* **46,** 352–356 (2014).

51. Wessel, J. *et al.* Low-frequency and rare exome chip variants associate with fasting glucose and type 2 diabetes susceptibility. *Nat. Commun.* **6,** 5897 (2015).

52. Richards, A. L. *et al.* Exome arrays capture polygenic rare variant contributions to schizophrenia. *Hum. Mol. Genet.* **25,** 1001–1007 (2016).

53. Auer, P. L. & Lettre, G. Rare variant association studies: considerations, challenges and opportunities. *Genome Med.* **7,** 16 (2015).

54. Wu, M. C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89,** 82–93 (2011).

55. Lee, S., Wu, M. C. & Lin, X. Optimal tests for rare variant effects in sequencing association studies. *Biostat. Oxf. Engl.* **13,** 762–775 (2012).

56. Jackson, T. M., Bittman, M. & Granowetter, L. Pediatric Malignant Bone Tumors: A Review and Update on Current Challenges, and Emerging Drug Targets. *Curr. Probl. Pediatr. Adolesc. Health Care* **46,** 213–228 (2016).

57. Esiashvili, N., Goodman, M. & Marcus, R. B. Changes in incidence and survival of Ewing sarcoma patients over the past 3 decades: Surveillance Epidemiology and End Results data. *J. Pediatr. Hematol. Oncol.* **30,** 425–430 (2008).

58. Biswas, B. & Bakhshi, S. Management of Ewing sarcoma family of tumors: Current scenario and unmet need. *World J. Orthop.* **7,** 527–538 (2016).

59. Jawad, M. U. *et al.* Ewing sarcoma demonstrates racial disparities in incidence-related and sex-related differences in outcome: an analysis of 1631 cases from the SEER database, 1973-2005. *Cancer* **115,** 3526–3536 (2009).

60. Worch, J., Matthay, K. K., Neuhaus, J., Goldsby, R. & DuBois, S. G. Ethnic and racial differences in patients with Ewing sarcoma. *Cancer* **116,** 983–988 (2010).

61. Arpaci, E. *et al.* Prognostic factors and clinical outcome of patients with Ewing's sarcoma family of tumors in adults: multicentric study of the Anatolian Society of Medical Oncology. *Med. Oncol. Northwood Lond. Engl.* **30,** 469 (2013).

62. Stahl, M. *et al.* Risk of recurrence and survival after relapse in patients with Ewing sarcoma. *Pediatr. Blood Cancer* **57,** 549–553 (2011).

63. Llombart Bosch, A., Machado, I. & Lopez-Guerrero, J. A. Biomarkers in the Ewing sarcoma family of tumors. *Curr. Biomark. Find.* **Volume 4,** 81–92 (2014).

64. Shukla, N. *et al.* Biomarkers in Ewing Sarcoma: The Promise and Challenge of Personalized Medicine. A Report from the Children's Oncology Group. *Front. Oncol.* **3,** 141 (2013).

65. Delattre, O. *et al.* Gene fusion with an ETS DNA-binding domain caused by chromosome translocation in human tumours. *Nature* **359,** 162–165 (1992).

66. Zoubek, A. *et al.* Does expression of different EWS chimeric transcripts define clinically distinct risk groups of Ewing tumor patients? *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **14,** 1245–1251 (1996).

67. de Alava, E., Lozano, M. D., Patiño, A., Sierrasesúmaga, L. & Pardo-Mindán, F. J. Ewing family tumors: potential prognostic value of reverse-transcriptase polymerase chain reaction detection of minimal residual disease in peripheral blood samples. *Diagn. Mol. Pathol. Am. J. Surg. Pathol. Part B* **7,** 152–157 (1998).

68. van Doorninck, J. A. *et al.* Current treatment protocols have eliminated the prognostic advantage of type 1 fusions in Ewing sarcoma: a report from the Children's Oncology Group. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **28,** 1989–1994 (2010).

69. Tirode, F. *et al.* Genomic landscape of Ewing sarcoma defines an aggressive subtype with co-association of STAG2 and TP53 mutations. *Cancer Discov.* **4,** 1342–1353 (2014).

70. Brohl, A. S. *et al.* The genomic landscape of the Ewing Sarcoma family of tumors reveals recurrent STAG2 mutation. *PLoS Genet.* **10,** e1004475 (2014).

71. Dubois, S. G., Epling, C. L., Teague, J., Matthay, K. K. & Sinclair, E. Flow cytometric detection of Ewing sarcoma cells in peripheral blood and bone marrow. *Pediatr. Blood Cancer* **54,** 13–18 (2010).

72. Ash, S. *et al.* Excellent prognosis in a subset of patients with Ewing sarcoma identified at diagnosis by CD56 using flow cytometry. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **17,** 2900–2907 (2011).

73. Fuchs, B., Inwards, C. Y. & Janknecht, R. Vascular endothelial growth factor expression is up-regulated by EWS-ETS oncoproteins and Sp1 and may represent an independent predictor of survival in Ewing's sarcoma. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **10,** 1344–1353 (2004).

74. Berghuis, D. *et al.* The CXCR4-CXCL12 axis in Ewing sarcoma: promotion of tumor growth rather than metastatic disease. *Clin. Sarcoma Res.* **2,** 24 (2012).

75. Jiang, X. *et al.* CD133 expression in chemo-resistant Ewing sarcoma cells. *BMC Cancer* **10,** 116 (2010).

76. Asmane, I. *et al.* Insulin-like growth factor type 1 receptor (IGF-1R) exclusive nuclear staining: a predictive biomarker for IGF-1R monoclonal antibody (Ab) therapy in sarcomas. *Eur. J. Cancer Oxf. Engl. 1990* **48,** 3027–3035 (2012).

77. Summary of the European public assessment report (EPAR) for Xeloda (Capecitabine) Tablets Prescribing Information available at the European Medicines Agency (EMA). (http://www.ema.europa.eu/ema/) Updated on July 27, 2016.

78. Hartkamp, A., van Boxtel, A. J., Zonnenberg, B. A. & Witteveen, P. O. Totally implantable venous access devices: evaluation of complications and a prospective comparative study of two different port systems. *Neth. J. Med.* **57,** 215–223 (2000).

79. Liu, G., Franssen, E., Fitch, M. I. & Warner, E. Patient preferences for oral versus intravenous palliative chemotherapy. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **15,** 110–115 (1997).

80. Walko, C. M. & Lindley, C. Capecitabine: a review. *Clin. Ther.* **27,** 23–44 (2005).

81. Reigner, B., Blesch, K. & Weidekamm, E. Clinical pharmacokinetics of capecitabine. *Clin. Pharmacokinet.* **40,** 85–104 (2001).

82. Mader, R. M. *et al.* Penetration of capecitabine and its metabolites into malignant and healthy tissue of patients with advanced breast cancer. *Int. J. Clin. Pharmacol. Ther.* **40,** 571–572 (2002).

83. Rose, M. G., Farrell, M. P. & Schmitz, J. C. Thymidylate synthase: a critical target for cancer chemotherapy. *Clin. Colorectal Cancer* **1,** 220–229 (2002).

84. Saif, M. W., Katirtzoglou, N. A. & Syrigos, K. N. Capecitabine: an overview of the side effects and their management. *Anticancer. Drugs* **19,** 447–464 (2008).

85. McKendrick, J. & Coutsouvelis, J. Capecitabine: effective oral fluoropyrimidine chemotherapy. *Expert Opin. Pharmacother.* **6,** 1231–1239 (2005).

86. U.S. Department of Health and Human services, National Cancer Institute. Cancer therapy Evaluation program-Common terminology Criteria for Adverse Events (CTCAE)-version 4.0. 2010.

87. Milano, G. *et al.* Candidate mechanisms for capecitabine-related hand-foot syndrome. *Br. J. Clin. Pharmacol.* **66,** 88–95 (2008).

88. Stein, B. N. *et al.* Age and sex are independent predictors of 5-fluorouracil toxicity. Analysis of a large scale phase III trial. *Cancer* **75,** 11–17 (1995).

89. Cassidy, J. *et al.* First-line oral capecitabine therapy in metastatic colorectal cancer: a favorable safety profile compared with intravenous 5-fluorouracil/leucovorin. *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol.* **13,** 566–575 (2002).

90. Haller, D. G. *et al.* Potential regional differences for the tolerability profiles of fluoropyrimidines. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **26,** 2118–2123 (2008).

91. Ribelles, N. *et al.* A carboxylesterase 2 gene polymorphism as predictor of capecitabine on response and time to progression. *Curr. Drug Metab.* **9,** 336–343 (2008).

92. Loganayagam, A. *et al.* The contribution of deleterious DPYD gene sequence variants to fluoropyrimidine toxicity in British cancer patients. *Cancer Chemother. Pharmacol.* **65,** 403–406 (2010).

93. Caronia, D. *et al.* A polymorphism in the cytidine deaminase promoter predicts severe capecitabine-induced hand-foot syndrome. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **17,** 2006–2013 (2011).

94. Deenen, M. J. *et al.* Relationship between single nucleotide polymorphisms and haplotypes in DPYD and toxicity and efficacy of capecitabine in advanced colorectal cancer. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **17,** 3455–3468 (2011).

95. Loganayagam, A. *et al.* Pharmacogenetic variants in the DPYD, TYMS, CDA and MTHFR genes are clinically significant predictors of fluoropyrimidine toxicity. *Br. J. Cancer* **108,** 2505–2515 (2013).

96. van Huis-Tanja, L. H., Gelderblom, H., Punt, C. J. A. & Guchelaar, H.-J. MTHFR polymorphisms and capecitabine-induced toxicity in patients with metastatic colorectal cancer. *Pharmacogenet. Genomics* **23,** 208–218 (2013).

97. Rosmarin, D. *et al.* Genetic markers of toxicity from capecitabine and other fluorouracil-based regimens: investigation in the QUASAR2 study, systematic review, and meta-analysis. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **32,** 1031–1039 (2014).

98. Rosmarin, D. *et al.* A candidate gene study of capecitabine-related toxicity in colorectal cancer identifies new toxicity variants at DPYD and a putative role for ENOSF1 rather than TYMS. *Gut* **64,** 111–120 (2015).

99. García-González, X. *et al.* Variants in CDA and ABCB1 are predictors of capecitabine-related adverse reactions in colorectal cancer. *Oncotarget* **6,** 6422–6430 (2015).

100. Di Marco, A., Cassinelli, G. & Arcamone, F. The discovery of daunorubicin. *Cancer Treat. Rep.* **65 Suppl 4,** 3–8 (1981).

101. Tan, C., Tasaka, H., Yu, K. P., Murphy, M. L. & Karnofsky, D. A. Daunomycin, an antitumor antibiotic, in the treatment of neoplastic disease. Clinical evaluation with special reference to childhood leukemia. *Cancer* **20,** 333–353 (1967).

102. Di Marco, A., Gaetani, M. & Scarpinato, B. Adriamycin (NSC-123,127): a new antibiotic with antitumor activity. *Cancer Chemother. Rep.* **53,** 33–37 (1969).

103. Volkova, M. & Russell, R. Anthracycline cardiotoxicity: prevalence, pathogenesis and treatment. *Curr. Cardiol. Rev.* **7,** 214–220 (2011).

104. Raj, S., Franco, V. I. & Lipshultz, S. E. Anthracycline-induced cardiotoxicity: a review of pathophysiology, diagnosis, and treatment. *Curr. Treat. Options Cardiovasc. Med.* **16,** 315 (2014).

105. Mulrooney, D. A. *et al.* Cardiac outcomes in a cohort of adult survivors of childhood and adolescent cancer: retrospective analysis of the Childhood Cancer Survivor Study cohort. *BMJ* **339,** b4606 (2009).

106. Tukenova, M. *et al.* Role of cancer treatment in long-term overall and cardiovascular mortality after childhood cancer. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **28,** 1308–1315 (2010).

107. van Dalen, E. C., Raphaël, M. F., Caron, H. N. & Kremer, L. C. Treatment including anthracyclines versus treatment not including anthracyclines for childhood cancer. *Cochrane Database Syst. Rev.* CD006647 (2009). doi:10.1002/14651858.CD006647.pub2

108. Lipshultz, S. E. *et al.* Long-term cardiovascular toxicity in children, adolescents, and young adults who receive cancer therapy: pathophysiology, course, monitoring, management, prevention, and research directions: a scientific statement from the American Heart Association. *Circulation* **128,** 1927–1995 (2013).

109. Minotti, G., Menna, P., Salvatorelli, E., Cairo, G. & Gianni, L. Anthracyclines: molecular advances and pharmacologic developments in antitumor activity and cardiotoxicity. *Pharmacol. Rev.* **56,** 185–229 (2004).

110. Harake, D., Franco, V. I., Henkel, J. M., Miller, T. L. & Lipshultz, S. E. Cardiotoxicity in childhood cancer survivors: strategies for prevention and management. *Future Cardiol.* **8,** 647–670 (2012).

111. Jensen, B. C. & McLeod, H. L. Pharmacogenomics as a risk mitigation strategy for chemotherapeutic cardiotoxicity. *Pharmacogenomics* **14,** 205–213 (2013).

112. Lipshultz, S. E., Alvarez, J. A. & Scully, R. E. Anthracycline associated cardiotoxicity in survivors of childhood cancer. *Heart Br. Card. Soc.* **94,** 525–533 (2008).

113. Doroshow, J. H., Locker, G. Y. & Myers, C. E. Enzymatic defenses of the mouse heart against reactive oxygen metabolites: alterations produced by doxorubicin. *J. Clin. Invest.* **65,** 128–135 (1980).

114. Barry, E., Alvarez, J. A., Scully, R. E., Miller, T. L. & Lipshultz, S. E. Anthracycline-induced cardiotoxicity: course, pathophysiology, prevention and management. *Expert Opin. Pharmacother.* **8,** 1039–1058 (2007).

115. Lipshultz, S. E. *et al.* Chronic progressive cardiac dysfunction years after doxorubicin therapy for childhood acute lymphoblastic leukemia. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **23,** 2629–2636 (2005).

116. Rosa, G. M. *et al.* Update on cardiotoxicity of anti-cancer treatments. *Eur. J. Clin. Invest.* **46,** 264–284 (2016).

117. van der Pal, H. J. *et al.* High risk of symptomatic cardiac events in childhood cancer survivors. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **30,** 1429–1437 (2012).

118. P. Fumoleau 1 , * , H. Roché 2 , P. Kerbrat 3 , J. Bonneterre 4 , P. Romestaing 5 , P. Fargeot 1 , M. Namer 6 , A. Monnier 7 , P. Montcuquet 8 , M.-J. Goudier 9 , E. Luporsi 10 and On behalf of the French Adjuvant Study Group. Long-term cardiac toxicity after adjuvant epirubicin-based chemotherapy in early breast cancer: French Adjuvant Study Group Results.

119. Lipshultz, S. E. *et al.* Late cardiac effects of doxorubicin therapy for acute lymphoblastic leukemia in childhood. *N. Engl. J. Med.* **324,** 808–815 (1991).

120. Shakir, D. K. & Rasul, K. I. Chemotherapy induced cardiomyopathy: pathogenesis, monitoring and management. *J. Clin. Med. Res.* **1,** 8–12 (2009).

121. Lipshultz, S. E., Cochran, T. R., Franco, V. I. & Miller, T. L. Treatment-related cardiotoxicity in survivors of childhood cancer. *Nat. Rev. Clin. Oncol.* **10,** 697–710 (2013).

122. Vejpongsa, P. & Yeh, E. T. H. Prevention of anthracycline-induced cardiotoxicity: challenges and opportunities. *J. Am. Coll. Cardiol.* **64,** 938–945 (2014).

123. Trachtenberg, B. H. *et al.* Anthracycline-associated cardiotoxicity in survivors of childhood cancer. *Pediatr. Cardiol.* **32,** 342–353 (2011).

124. Wojnowski, L. *et al.* NAD(P)H oxidase and multidrug resistance protein genetic polymorphisms are associated with doxorubicin-induced cardiotoxicity. *Circulation* **112,** 3754–3762 (2005).

125. Blanco, J. G. *et al.* Genetic polymorphisms in the carbonyl reductase 3 gene CBR3 and the NAD(P)H:quinone oxidoreductase 1 gene NQO1 in patients who developed anthracycline-related congestive heart failure after childhood cancer. *Cancer* **112,** 2789–2795 (2008).

126. Rajić, V. *et al.* Influence of the polymorphism in candidate genes on late cardiac damage in patients treated due to acute leukemia in childhood. *Leuk. Lymphoma* **50,** 1693–1698 (2009).

127. Blanco, J. G. *et al.* Anthracycline-related cardiomyopathy after childhood cancer: role of polymorphisms in carbonyl reductase genes--a report from the Children's Oncology Group. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **30,** 1415–1421 (2012).

128. Lipshultz, S. E. *et al.* Impact of hemochromatosis gene mutations on cardiac status in doxorubicin-treated survivors of childhood high-risk leukemia. *Cancer* **119,** 3555–3562 (2013).

129. Semsei, A. F. *et al.* ABCC1 polymorphisms in anthracycline-induced cardiotoxicity in childhood acute lymphoblastic leukaemia. *Cell Biol. Int.* **36,** 79–86 (2012).

130. Cascales, A. *et al.* Clinical and genetic determinants of anthracycline-induced cardiac iron accumulation. *Int. J. Cardiol.* **154,** 282–286 (2012).

131. Cascales, A. *et al.* Association of anthracycline-related cardiac histological lesions with NADPH oxidase functional polymorphisms. *The Oncologist* **18,** 446–453 (2013).

132. Volkan-Salanci, B. *et al.* The relationship between changes in functional cardiac parameters following anthracycline therapy and carbonyl reductase 3 and glutathione S transferase Pi polymorphisms. *J. Chemother. Florence Italy* **24,** 285–291 (2012).

133. Visscher, H. *et al.* Pharmacogenomic prediction of anthracycline-induced cardiotoxicity in children. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **30,** 1422–1428 (2012).

134. Visscher, H. *et al.* Validation of variants in SLC28A3 and UGT1A6 as genetic markers predictive of anthracycline-induced cardiotoxicity in children. *Pediatr. Blood Cancer* **60,** 1375–1381 (2013).

135. Lubieniecka, J. M. *et al.* A discovery study of daunorubicin induced cardiotoxicity in a sample of acute myeloid leukemia patients prioritizes P450 oxidoreductase polymorphisms as a potential risk factor. *Front. Genet.* **4,** 231 (2013).

136. Wang, X. *et al.* Hyaluronan synthase 3 variant and anthracycline-related cardiomyopathy: a report from the children's oncology group. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **32,** 647–653 (2014).

137. Visscher, H. *et al.* Genetic variants in SLC22A17 and SLC22A7 are associated with anthracycline-induced cardiotoxicity in children. *Pharmacogenomics* **16,** 1065–1076 (2015).

138. Krajinovic, M. *et al.* Polymorphisms of ABCC5 and NOS3 genes influence doxorubicin cardiotoxicity in survivors of childhood acute lymphoblastic leukemia. *Pharmacogenomics J.* (2015). doi:10.1038/tpj.2015.63

139. Aminkeng, F. *et al.* A coding variant in RARG confers susceptibility to anthracycline-induced cardiotoxicity in childhood cancer. *Nat. Genet.* **47,** 1079–1084 (2015).

140. The Pharmacogenomics Knowledge database: htpps://www.pharmgkb.com.

141. Caronia, D. *et al.* Effect of ABCB1 and ABCC3 polymorphisms on osteosarcoma survival after chemotherapy: a pharmacogenetic study. *PloS One* **6,** e26091 (2011).

142. http://bioinfo.cipf.es/pupasuite/www/index.jsp.

143. http://www.broad.mit.edu/mpg/haploview.

144. http://www.illumina.com/support/documentation/VeraCode.ilmn.

145. Wunder, J. S. *et al.* The histological response to chemotherapy as a predictor of the oncological outcome of operative treatment of Ewing sarcoma. *J. Bone Joint Surg. Am.* **80,** 1020–1033 (1998).

146. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).

147. PLINK software: http://pngu.mgh.harvard.edu/~purcell/plink/index.shtml.

148. ENCODE Project Consortium. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.* **9,** e1001046 (2011).

149. UCSC Genome Browser: http://genome.ucsc.edu/index.html.

150. Haploreg web interface: http://www.broadinstitute.org/mammals/haploreg/haploreg.php.

151. Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* **40,** D930-934 (2012).

152. Gurwitz, D. & McLeod, H. L. Genome-wide studies in pharmacogenomics: harnessing the power of extreme phenotypes. *Pharmacogenomics* **14,** 337–339 (2013).

153. http://www.illumina.com/support/documentation/infinium_assay.ilmn.

154. Gabriel, S., Ziaugra, L. & Tabbaa, D. SNP genotyping using the Sequenom MassARRAY iPLEX platform. *Curr. Protoc. Hum. Genet.* **Chapter 2,** Unit 2.12 (2009).

155. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38,** 904–909 (2006).

156. Stephens, M. & Donnelly, P. A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am. J. Hum. Genet.* **73,** 1162–1169 (2003).

157. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5,** e1000529 (2009).

158. Hennings, H. *et al.* Calcium regulation of growth and differentiation of mouse epidermal cells in culture. *Cell* **19,** 245–254 (1980).

159. van de Werken, H. J. G. *et al.* 4C technology: protocols and data analysis. *Methods Enzymol.* **513,** 89–112 (2012).

160. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods San Diego Calif* **25,** 402–408 (2001).

161. Varghese, F., Bukhari, A. B., Malhotra, R. & De, A. IHC Profiler: an open source plugin for the quantitative evaluation and automated scoring of immunohistochemistry images of human tissue samples. *PloS One* **9,** e96801 (2014).

162. Maae, E. *et al.* Estimation of immunohistochemical expression of VEGF in ductal carcinomas of the breast. *J. Histochem. Cytochem. Off. J. Histochem. Soc.* **59,** 750–760 (2011).

163. Goldstein, J. I. *et al.* zCall: a rare variant caller for array-based genotyping: genetics and population analysis. *Bioinforma. Oxf. Engl.* **28,** 2543–2545 (2012).

164. http://support.illumina.com/content/illumina-support/us/en/array/array_kits/infinium_humanexome_beadchip_kit/downloads.html.

165. Yosef Hochberg, Y. B. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* **57,** 289–300 (1995).

166. Lee, S. *et al.* Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet.* **91,** 224–237 (2012).

167. Aulchenko, Y. S., Ripke, S., Isaacs, A. & van Duijn, C. M. GenABEL: an R library for genome-wide association analysis. *Bioinforma. Oxf. Engl.* **23,** 1294–1296 (2007).

168. Dennis, G. *et al.* DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.* **4,** P3 (2003).

169. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4,** 44–57 (2009).

170. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25,** 25–29 (2000).

171. Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Res.* **43,** D1049-1056 (2015).

172. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28,** 27–30 (2000).

173. Kanehisa, M. *et al.* Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.* **42,** D199-205 (2014).

174. Croft, D. *et al.* The Reactome pathway knowledgebase. *Nucleic Acids Res.* **42,** D472-477 (2014).

175. Fabregat, A. *et al.* The Reactome pathway Knowledgebase. *Nucleic Acids Res.* **44,** D481-487 (2016).

176. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4,** 1073–1081 (2009).

177. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7,** 248–249 (2010).

178. Li, B. *et al.* Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinforma. Oxf. Engl.* **25,** 2744–2750 (2009).

179. Calabrese, R., Capriotti, E., Fariselli, P., Martelli, P. L. & Casadio, R. Functional annotations improve the predictive score of human disease-related mutations in proteins. *Hum. Mutat.* **30,** 1237–1244 (2009).

180. Niroula, A., Urolagin, S. & Vihinen, M. PON-P2: prediction method for fast and reliable identification of harmful variants. *PloS One* **10,** e0117380 (2015).

181. Bendl, J. *et al.* PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations. *PLoS Comput. Biol.* **10,** e1003440 (2014).

182. Lee, P. H. & Shatkay, H. F-SNP: computationally predicted functional SNPs for disease association studies. *Nucleic Acids Res.* **36,** D820-824 (2008).

183. Halbleib, J. M. & Nelson, W. J. Cadherins in development: cell adhesion, sorting, and tissue morphogenesis. *Genes Dev.* **20,** 3199–3214 (2006).

184. El-Amraoui, A. & Petit, C. Cadherin defects in inherited human diseases. *Prog. Mol. Biol. Transl. Sci.* **116,** 361–384 (2013).

185. Dua-Awereh, M. B., Shimomura, Y., Kraemer, L., Wajid, M. & Christiano, A. M. Mutations in the desmoglein 1 gene in five Pakistani families with striate palmoplantar keratoderma. *J. Dermatol. Sci.* **53,** 192–197 (2009).

186. Rickman, L. *et al.* N-terminal deletion in a desmosomal cadherin causes the autosomal dominant skin disease striate palmoplantar keratoderma. *Hum. Mol. Genet.* **8,** 971–976 (1999).

187. Hennies, H. C., Küster, W., Mischke, D. & Reis, A. Localization of a locus for the striated form of palmoplantar keratoderma to chromosome 18q near the desmosomal cadherin gene cluster. *Hum. Mol. Genet.* **4,** 1015–1020 (1995).

188. Gibcus, J. H. & Dekker, J. The hierarchy of the 3D genome. *Mol. Cell* **49,** 773–782 (2013).

189. Handoko, L. *et al.* CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat. Genet.* **43,** 630–638 (2011).

190. DeMare, L. E. *et al.* The genomic landscape of cohesin-associated chromatin interactions. *Genome Res.* **23,** 1224–1234 (2013).

191. Candi, E., Schmidt, R. & Melino, G. The cornified envelope: a model of cell death in the skin. *Nat. Rev. Mol. Cell Biol.* **6,** 328–340 (2005).

192. Bartlett, K. & Eaton, S. Mitochondrial beta-oxidation. *Eur. J. Biochem.* **271,** 462–469 (2004).

193. Venkatesan, B. *et al.* WNT1-inducible signaling pathway protein-1 activates diverse cell survival pathways and blocks doxorubicin-induced cardiomyocyte death. *Cell. Signal.* **22,** 809–820 (2010).

194. Wild, C.P., S., B. International Agency for Research on Cancer, WHO. (2014) World Cancer Report 2014 [Online]. Available from: http://www.thehealthwell.info/node/725845 [Accessed: 10th November 2016].

195. Cascorbi, I. & Haenisch, S. Pharmacogenetics of ATP-binding cassette transporters and clinical implications. *Methods Mol. Biol. Clifton NJ* **596,** 95–121 (2010).

196. Huang, Y. Pharmacogenetics/genomics of membrane transporters in cancer chemotherapy. *Cancer Metastasis Rev.* **26,** 183–201 (2007).

197. Le Saux, O. *et al.* The molecular and physiological roles of ABCC6: more than meets the eye. *Front. Genet.* **3,** 289 (2012).

198. Faber, K. N., Müller, M. & Jansen, P. L. M. Drug transport proteins in the liver. *Adv. Drug Deliv. Rev.* **55,** 107–124 (2003).

199. Marin, J. J. G. Plasma membrane transporters in modern liver pharmacology. *Scientifica* **2012,** 428139 (2012).

200. Madon, J., Hagenbuch, B., Landmann, L., Meier, P. J. & Stieger, B. Transport function and hepatocellular localization of mrp6 in rat liver. *Mol. Pharmacol.* **57,** 634–641 (2000).

201. Sekine, T., Miyazaki, H. & Endou, H. Molecular physiology of renal organic anion transporters. *Am. J. Physiol. Renal Physiol.* **290,** F251-261 (2006).

202. Belinsky, M. G., Chen, Z.-S., Shchaveleva, I., Zeng, H. & Kruh, G. D. Characterization of the drug resistance and transport properties of multidrug resistance protein 6 (MRP6, ABCC6). *Cancer Res.* **62,** 6172–6177 (2002).

203. Kruh, G. D. *et al.* MRP subfamily transporters and resistance to anticancer agents. *J. Bioenerg. Biomembr.* **33,** 493–501 (2001).

204. Svoboda, L. K. *et al.* Overexpression of HOX genes is prevalent in Ewing sarcoma and is associated with altered epigenetic regulation of developmental transcription programs. *Epigenetics* **9,** 1613–1625 (2015).

205. Ekhart, C., Rodenhuis, S., Smits, P. H. M., Beijnen, J. H. & Huitema, A. D. R. An overview of the relations between polymorphisms in drug metabolising enzymes and drug transporters and survival after cancer drug treatment. *Cancer Treat. Rev.* **35,** 18–31 (2009).

206. Chen, Z.-S. & Tiwari, A. K. Multidrug resistance proteins (MRPs/ABCCs) in cancer chemotherapy and genetic diseases. *FEBS J.* **278,** 3226–3245 (2011).

207. Zhang, X., Bailey, S. D. & Lupien, M. Laying a solid foundation for Manhattan-–'setting the functional basis for the post-GWAS era'. *Trends Genet. TIG* **30,** 140–149 (2014).

208. Symmons, O. *et al.* Functional and topological characteristics of mammalian regulatory domains. *Genome Res.* **24,** 390–400 (2014).

209. Pombo, A. & Dillon, N. Three-dimensional genome architecture: players and mechanisms. *Nat. Rev. Mol. Cell Biol.* **16,** 245–257 (2015).

210. Chung, I.-M. *et al.* Making Sense of the Tangle: Insights into Chromatin Folding and Gene Regulation. *Genes* **7,** (2016).

211. Ong, C.-T. & Corces, V. G. CTCF: an architectural protein bridging genome topology and function. *Nat. Rev. Genet.* **15,** 234–246 (2014).

212. Nasmyth, K. & Haering, C. H. Cohesin: its roles and mechanisms. *Annu. Rev. Genet.* **43,** 525–558 (2009).

213. Merkenschlager, M. & Odom, D. T. CTCF and cohesin: linking gene regulatory elements with their targets. *Cell* **152,** 1285–1297 (2013).

214. Wendt, K. S. *et al.* Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature* **451,** 796–801 (2008).

215. Chan, L. M. S., Lowes, S. & Hirst, B. H. The ABCs of drug transport in intestine and liver: efflux proteins limiting drug absorption and bioavailability. *Eur. J. Pharm. Sci. Off. J. Eur. Fed. Pharm. Sci.* **21,** 25–51 (2004).

216. Tiwari, A. K., Sodani, K., Dai, C.-L., Ashby, C. R. & Chen, Z.-S. Revisiting the ABCs of multidrug resistance in cancer chemotherapy. *Curr. Pharm. Biotechnol.* **12,** 570–594 (2011).

217. Leschziner, G. D., Andrew, T., Pirmohamed, M. & Johnson, M. R. ABCB1 genotype and PGP expression, function and therapeutic drug response: a critical review and recommendations for future research. *Pharmacogenomics J.* **7,** 154–179 (2007).

218. Perri, T. *et al.* Effect of P-glycoprotein expression on outcome in the Ewing family of tumors. *Pediatr. Hematol. Oncol.* **18,** 325–334 (2001).

219. Roundhill, E. & Burchill, S. Membrane expression of MRP-1, but not MRP-1 splicing or Pgp expression, predicts survival in patients with ESFT. *Br. J. Cancer* **109,** 195–206 (2013).

220. Roessner, A. *et al.* Prognostic implication of immunodetection of P glycoprotein in Ewing's sarcoma. *J. Cancer Res. Clin. Oncol.* **119,** 185–189 (1993).

221. Fung, K. L. & Gottesman, M. M. A synonymous polymorphism in a common MDR1 (ABCB1) haplotype shapes protein function. *Biochim. Biophys. Acta* **1794,** 860–871 (2009).

222. Bledsoe, K. L. *et al.* RUNX3 Facilitates Growth of Ewing Sarcoma Cells. *J. Cell. Physiol.* **229,** 2049–2056 (2014).

223. Daily, E. B. & Aquilante, C. L. Cytochrome P450 2C8 pharmacogenetics: a review of clinical studies. *Pharmacogenomics* **10,** 1489–1510 (2009).

224. van Schaik, R. H. N. Cancer treatment and pharmacogenetics of cytochrome P450 enzymes. *Invest. New Drugs* **23,** 513–522 (2005).

225. Totah, R. A. & Rettie, A. E. Cytochrome P450 2C8: substrates, inhibitors, pharmacogenetics, and clinical relevance. *Clin. Pharmacol. Ther.* **77,** 341–352 (2005).

226. Chang, T. K., Weber, G. F., Crespi, C. L. & Waxman, D. J. Differential activation of cyclophosphamide and ifosphamide by cytochromes P-450 2B and 3A in human liver microsomes. *Cancer Res.* **53,** 5629–5637 (1993).

227. Gardiner, S. J. & Begg, E. J. Pharmacogenetics, drug-metabolizing enzymes, and clinical practice. *Pharmacol. Rev.* **58,** 521–590 (2006).

228. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337,** 1190–1195 (2012).

229. Chen, C.-Y., Chang, I.-S., Hsiung, C. A. & Wasserman, W. W. On the identification of potential regulatory variants within genome wide association candidate SNP sets. *BMC Med. Genomics* **7,** 34 (2014).

230. Nicolae, D. L. *et al.* Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* **6,** e1000888 (2010).

231. Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S. & Snyder, M. Linking disease associations with regulatory information in the human genome. *Genome Res.* **22,** 1748–1759 (2012).

232. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473,** 43–49 (2011).

233. Tang, Z. *et al.* CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. *Cell* **163,** 1611–1627 (2015).

234. Ding, Z. *et al.* Quantitative genetics of CTCF binding reveal local sequence effects and different modes of X-chromosome association. *PLoS Genet.* **10,** e1004798 (2014).

235. Levo, M. *et al.* Unraveling determinants of transcription factor binding outside the core binding site. *Genome Res.* **25,** 1018–1029 (2015).

236. Dror, I., Golan, T., Levy, C., Rohs, R. & Mandel-Gutfreund, Y. A widespread role of the motif environment in transcription factor binding across diverse protein families. *Genome Res.* **25,** 1268–1280 (2015).

237. Takeichi, M. Morphogenetic roles of classic cadherins. *Curr. Opin. Cell Biol.* **7,** 619–627 (1995).

238. Simpson, C. L., Patel, D. M. & Green, K. J. Deconstructing the skin: cytoarchitectural determinants of epidermal morphogenesis. *Nat. Rev. Mol. Cell Biol.* **12,** 565–580 (2011).

239. Furukawa, F. *et al.* Roles of E- and P-cadherin in the human skin. *Microsc. Res. Tech.* **38,** 343–352 (1997).

240. Bikle, D. D., Xie, Z. & Tu, C.-L. Calcium regulation of keratinocyte differentiation. *Expert Rev. Endocrinol. Metab.* **7,** 461–472 (2012).

241. Rosenberg, P. *et al.* A potential role of R-cadherin in striated muscle formation. *Dev. Biol.* **187,** 55–70 (1997).

242. Redies, C., Engelhart, K. & Takeichi, M. Differential expression of N- and R-cadherin in functional neuronal systems and other structures of the developing chicken brain. *J. Comp. Neurol.* **333,** 398–416 (1993).

243. Steinert, P. M. & Marekov, L. N. Direct evidence that involucrin is a major early isopeptide cross-linked component of the keratinocyte cornified cell envelope. *J. Biol. Chem.* **272,** 2021–2030 (1997).

244. Nemes, Z., Marekov, L. N., Fésüs, L. & Steinert, P. M. A novel function for transglutaminase 1: Attachment of long-chain ω-hydroxyceramides to involucrin by ester bond formation. *Proc. Natl. Acad. Sci. U. S. A.* **96,** 8402–8407 (1999).

245. Koch, P. J. *et al.* Lessons from loricrin-deficient mice: compensatory mechanisms maintaining skin barrier function in the absence of a major cornified envelope protein. *J. Cell Biol.* **151,** 389–400 (2000).

246. Djian, P., Easley, K. & Green, H. Targeted ablation of the murine involucrin gene. *J. Cell Biol.* **151,** 381–388 (2000).

247. Aho, S. *et al.* Periplakin gene targeting reveals a constituent of the cornified cell envelope dispensable for normal mouse development. *Mol. Cell. Biol.* **24,** 6410–6418 (2004).

248. Määttä, A., DiColandrea, T., Groot, K. & Watt, F. M. Gene targeting of envoplakin, a cytoskeletal linker protein and precursor of the epidermal cornified envelope. *Mol. Cell. Biol.* **21,** 7047–7053 (2001).

249. Sevilla, L. M. *et al.* Mice deficient in involucrin, envoplakin, and periplakin have a defective epidermal barrier. *J. Cell Biol.* **179,** 1599–1612 (2007).

250. Sugiura, H. *et al.* Large-scale DNA microarray analysis of atopic skin lesions shows overexpression of an epidermal differentiation gene cluster in the alternative pathway and lack of protective gene expression in the cornified envelope. *Br. J. Dermatol.* **152,** 146–149 (2005).

251. Guttman-Yassky, E. *et al.* Broad defects in epidermal cornification in atopic dermatitis identified through genomic analysis. *J. Allergy Clin. Immunol.* **124,** 1235–1244.e58 (2009).

252. Chen, J.-Q. *et al.* Regulation of involucrin in psoriatic epidermal keratinocytes: the roles of ERK1/2 and GSK-3β. *Cell Biochem. Biophys.* **66,** 523–528 (2013).

253. Peña-Penabad, C. *et al.* Altered expression of immunoreactive involucrin in lamellar ichthyosis. *Eur. J. Dermatol. EJD* **9,** 197–201 (1999).

254. Li, B. & Leal, S. M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.* **83,** 311–321 (2008).

255. Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.* **95,** 5–23 (2014).

256. Pierce, K. L., Premont, R. T. & Lefkowitz, R. J. Seven-transmembrane receptors. *Nat. Rev. Mol. Cell Biol.* **3,** 639–650 (2002).

257. Sun, Y. V. *et al.* Application of machine learning algorithms to predict coronary artery calcification with a sibship-based design. *Genet. Epidemiol.* **32,** 350–360 (2008).

258. Min, K.-D. *et al.* Identification of genes related to heart failure using global gene expression profiling of human failing myocardium. *Biochem. Biophys. Res. Commun.* **393,** 55–60 (2010).

259. Ronkainen, V.-P. *et al.* Hypoxia-inducible factor 1-induced G protein-coupled receptor 35 expression is an early marker of progressive cardiac remodelling. *Cardiovasc. Res.* **101,** 69–77 (2014).

260. Wang, J. *et al.* Kynurenic acid as a ligand for orphan G protein-coupled receptor GPR35. *J. Biol. Chem.* **281,** 22021–22028 (2006).

261. Geft, I. L. *et al.* Intermittent brief periods of ischemia have a cumulative effect and may cause myocardial necrosis. *Circulation* **66,** 1150–1153 (1982).

262. Watanabe, Y., Kusuoka, H., Fukuchi, K., Fujiwara, T. & Nishimura, T. Contribution of hypoxia to the development of cardiomyopathy in hamsters. *Cardiovasc. Res.* **35,** 217–222 (1997).

263. Offermanns, S., Mancino, V., Revel, J. P. & Simon, M. I. Vascular system defects and impaired cell chemokinesis as a result of Galpha13 deficiency. *Science* **275,** 533–536 (1997).

264. Ruppel, K. M. *et al.* Essential role for Galpha13 in endothelial cells during embryonic development. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 8281–8286 (2005).

265. Li, Y. & Anand-Srivastava, M. B. Inactivation of enhanced expression of G(i) proteins by pertussis toxin attenuates the development of high blood pressure in spontaneously hypertensive rats. *Circ. Res.* **91,** 247–254 (2002).

266. Maruyama, Y. *et al.* Galpha(12/13) mediates alpha(1)-adrenergic receptor-induced cardiac hypertrophy. *Circ. Res.* **91,** 961–969 (2002).

267. Schober, A. & Siess, W. Lysophosphatidic acid in atherosclerotic diseases. *Br. J. Pharmacol.* **167,** 465–482 (2012).

268. Mizutani, K. *et al.* Kynureninase is a novel candidate gene for hypertension in spontaneously hypertensive rats. *Hypertens. Res. Off. J. Jpn. Soc. Hypertens.* **25,** 135–140 (2002).

269. Hamilton, M. A., Stevenson, L. W., Luu, M. & Walden, J. A. Altered thyroid hormone metabolism in advanced heart failure. *J. Am. Coll. Cardiol.* **16,** 91–95 (1990).

270. Friberg, L., Werner, S., Eggertsen, G. & Ahnve, S. Rapid down-regulation of thyroid hormones in acute myocardial infarction: is it cardioprotective in patients with angina? *Arch. Intern. Med.* **162,** 1388–1394 (2002).

271. Ishizaki, K. *et al.* The mitochondrial electron transfer flavoprotein complex is essential for survival of Arabidopsis in extended darkness. *Plant J. Cell Mol. Biol.* **47,** 751–760 (2006).

272. Abdel-aleem, S., el-Merzabani, M. M., Sayed-Ahmed, M., Taylor, D. A. & Lowe, J. E. Acute and chronic effects of adriamycin on fatty acid oxidation in isolated cardiac myocytes. *J. Mol. Cell. Cardiol.* **29,** 789–797 (1997).

273. Kawasaki, N., Lee, J. D., Shimizu, H. & Ueda, T. Long-term l-carnitine treatment prolongs the survival in rats with adriamycin-induced heart failure. *J. Card. Fail.* **2,** 293–299 (1996).

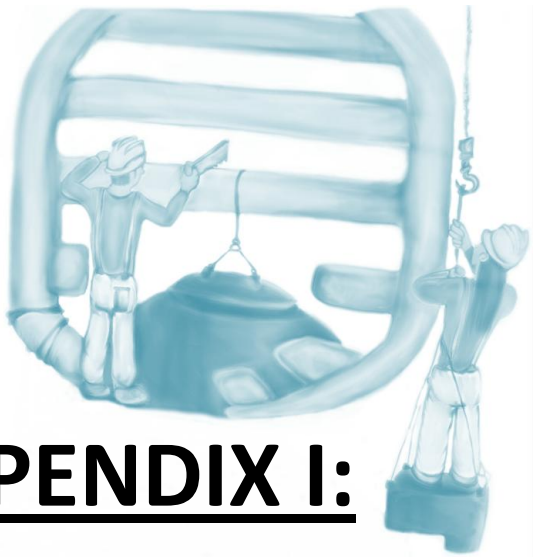
274. Tokarska-Schlattner, M., Zaugg, M., Zuppinger, C., Wallimann, T. & Schlattner, U. New insights into doxorubicin-induced cardiotoxicity: the critical role of cellular energetics. *J. Mol. Cell. Cardiol.* **41,** 389–405 (2006).

275. Kumar, S. N., Konorev, E. A., Aggarwal, D. & Kalyanaraman, B. Analysis of proteome changes in doxorubicin-treated adult rat cardiomyocyte. *J. Proteomics* **74,** 683–697 (2011).

276. Chen, Y. *et al.* Redox proteomic identification of oxidized cardiac proteins in adriamycin-treated mice. *Free Radic. Biol. Med.* **41,** 1470–1477 (2006).

277. Venkatachalam, K. *et al.* WISP1, a pro-mitogenic, pro-survival factor, mediates tumor necrosis factor-alpha (TNF-alpha)-stimulated cardiac fibroblast proliferation but inhibits TNF-alpha-induced cardiomyocyte death. *J. Biol. Chem.* **284,** 14414–14427 (2009).

278. Colston, J. T. *et al.* Wnt-induced secreted protein-1 is a prohypertrophic and profibrotic growth factor. *Am. J. Physiol. Heart Circ. Physiol.* **293,** H1839-1846 (2007).



APPENDIX I:

publications derived from the thesis

- **Ruiz-Pinto S**, Pita G, Patiño-García A, García-Miguel P, Alonso J, Pérez-Martínez A, Sastre A, Gómez-Mariano G, Lissat A, Scotlandi K, Serra M, Ladenstein R, Lapouble E, Pierron G, Kontny U, Picci P, Kovar H, Delattre O, González-Neira A. Identification of genetic variants in pharmacokinetic genes associated with Ewing Sarcoma treatment outcome. *Ann Oncol Off J Eur Soc Med Onco*l. **9**:1788–93 (2016).

- **Ruiz-Pinto S*,** Pita G*, Martín M*, Cuadrado A, Shahbazi MN, Caronia D, Kojic A, Moreno LT, de la Torre-Montero JC, Lozano M, López-Fernández LA, Ribelles N, García-Saenz JA, Alba E, Milne RL, Losada A, Pérez-Moreno M, Benítez J, González-Neira A. Cis-acting regulatory variants at the *CDH4* gene locus reveal a novel mechanism of susceptibility to capecitabine–induced hand-foot syndrome. Under review in Journal of Clinical Oncology.

- Ruiz-Pinto S, Pita G, Patiño-García A, García-Miguel P, Alonso J, Pérez-Martínez A, Cartón AJ, Gutiérrez-Larraya F, Alonso MR, Barnes, DR, Dennis J, Michailidou K, Gómez-Santos C, Thompson DJ, Easton DF, Benítez J, González-Neira A. Exome array analysis identifies *GPR35* as a novel susceptibility gene for anthracycline-induced cardiotoxicity in childhood cancer. Under review in Pharmacogenetics and Genomics.

- **Ruiz-Pinto S**, Pita G, Martin M, Alonso-Gordoa T, Vulsteke C, Alonso MR, Herráez B, Cartón AJ, Peuteman G, García-Miguel P, Alonso J, Pérez-Martínez A, Gutiérrez-Larraya F, Lambrechts D, Wildiers H, García-Sáenz JA, Patiño-García A, González-Neira A. Exome array analysis identifies *ETFB* and *WISP1* as novel susceptibility genes for anthracycline-induced cardiotoxicity in cancer patients. Under review in Annals of Oncology.

Annals of Oncology

Annals of Oncology 27: 1788–1793, 2016
doi:10.1093/annonc/mdw234
Published online 10 June 2016

Identification of genetic variants in pharmacokinetic genes associated with Ewing Sarcoma treatment outcome

S. Ruiz-Pinto[1], G. Pita[1], A. Patiño-García[2], P. García-Miguel[3], J. Alonso[4], A. Pérez-Martínez[3], A. Sastre[3], G. Gómez-Mariano[4], A. Lissat[5], K. Scotlandi[6], M. Serra[6], R. Ladenstein[7], E. Lapouble[8], G. Pierron[8], U. Kontny[9], P. Picci[6], H. Kovar[7], O. Delattre[10] & A. González-Neira[1]*

[1]*Human Genotyping Unit-CeGen, Human Cancer Genetics Programme, Spanish National Cancer Research Centre (CNIO), Madrid, Spain;* [2]*Clinical Genetics Unit, University Clinic of Navarra (CUN), Pamplona, Spain;* [3]*Department of Pediatric Hemato-Oncology, Hospital Universitario La Paz, Madrid, Spain;* [4]*Pediatric Solid Tumor Laboratory, Human Genetic Department, Research Institute of Rare Diseases, Instituto de Salud Carlos III, Majadahonda, Madrid, Spain;* [5]*Department of Pediatrics, Division of Oncology and Hematology, Charité Universitaetsmedizin, Berlin, Germany;* [6]*Experimental Oncology Laboratory, Istituto Ortopedico Rizzoli, Bologna, Italy;* [7]*Department of Pediatrics, Children's Cancer Research Institute, St Anna Kinderkrebsforschung e.V., Medical University, Vienna, Austria;* [8]*Somatic Genetics Unit, Institut Curie, Paris, France;* [9]*Division of Paediatric Haematology, Oncology and Stem Cell Transplantation, Department of Paediatrics and Adolescent Medicine, University Medical Centre, Aachen, Germany;* [10]*Inserm U830, Centre de Recherche, Institut Curie, Paris, France*

Received 2 February 2016; revised 21 April 2016; accepted 30 May 2016

Background: Despite the effectiveness of current treatment protocols for Ewing sarcoma (ES), many patients still experience relapse, and survival following recurrence is <15%. We aimed to identify genetic variants that predict treatment outcome in children diagnosed with ES.

Patients and methods: We carried out a pharmacogenetic study of 384 single-nucleotide polymorphisms (SNPs) in 24 key transport or metabolism genes relevant to drugs used to treat in pediatric patients (<30 years) with histologically confirmed ES. We studied the association of genotypes with tumor response and overall survival (OS) in a discovery cohort of 106 Spanish children, with replication in a second cohort of 389 pediatric patients from across Europe.

Results: We identified associations with OS ($P < 0.05$) for three SNPs in the Spanish cohort that were replicated in the European cohort. The strongest association observed was with rs7190447, located in the ATP-binding cassette subfamily C member 6 (*ABCC6*) gene [discovery: hazard ratio (HR) = 14.30, 95% confidence interval (CI) = 1.53–134, $P = 0.020$; replication: HR = 9.28, 95% CI = 2.20–39.2, $P = 0.0024$] and its correlated SNP rs7192303, which was predicted to have a plausible regulatory function. We also replicated associations with rs4148737 in the ATP-binding cassette subfamily B member 1 (*ABCB1*) gene (discovery: HR = 2.96, 95% CI = 1.08–8.10, $P = 0.034$; replication: HR = 1.60, 95% CI = 1.05–2.44, $P = 0.029$), which we have previously found to be associated with poorer OS in pediatric osteosarcoma patients, and rs11188147 in cytochrome P450 family 2 subfamily C member 8 gene *(CYP2C8)* (discovery : HR = 2.49, 95% CI = 1.06–5.87, $P = 0.037$; replication: HR = 1.77, 95% CI = 1.06–2.96, $P = 0.030$), an enzyme involved in the oxidative metabolism of the ES chemotherapeutic agents cyclophosphamide and ifosfamide. None of the associations with tumor response were replicated.

Conclusion: Using an integrated pathway-based approach, we identified polymorphisms in *ABCC6*, *ABCB1* and *CYP2C8* associated with OS. These associations were replicated in a large independent cohort, highlighting the importance of pharmacokinetic genes as prognostic markers in ES.

Key words: Ewing sarcoma, polymorphisms, pharmacokinetic genes, prognostic, pathway-based approach

introduction

Ewing sarcoma (ES) is relatively uncommon, despite being the second most frequent primary malignant bone tumor in children and adolescents, after osteosarcoma. It accounts for only 2% of all childhood cancers, with an annual incidence of ~3 cases per million [1]. Interindividual variability in drug efficacy and toxicity, resulting in unpredictable patient response, is of particular concern for chemotherapeutic drugs because these agents have a narrow therapeutic window and must be given at optimal doses [2]. In particular in ES, 30%–40% of patients with a localized primary tumor and 60%–80% of patients with disseminated disease experience relapse after treatment and have a dismal prognosis, with a likelihood of long-term survival after recurrence lower than 15% [3, 4]. We hypothesized that genetic

*Correspondence to: Dr. Anna Gonzalez-Neira, Human Genotyping Unit-CeGen, Human Cancer Genetics Programme, Spanish National Cancer Centre, Melchor Fernández Almagro 3, Madrid 28029, Spain. Tel: +34-91-2246974; Fax: +34-91-2246923; E-mail: agonzalez@cnio.es

variants in genes encoding drug transporters, drug-metabolizing enzymes and/or drug targets might explain much of this variability in drug response [5, 6]. In order to identify prognostic and predictive markers for ES, we conducted a pharmacogenetic study genotyping 384 single-nucleotide polymorphisms (SNPs) in 24 genes involved in the absorption, distribution, metabolism and elimination of chemotherapeutic drugs used to treat ES using a discovery cohort from Spain ($n = 106$), with replication in an independent European cohort ($n = 389$).

materials and methods

patients

Eligible patients had histologically confirmed ES diagnosed before age 30 years. The discovery cohort consisted of 106 Spanish ES pediatric patients recruited between 1993 and 2012 at the University Hospital La Paz and University Hospital Niño Jesús in Madrid and at the University Clinic of Navarra in Pamplona. The replication cohort consisted of 389 ES pediatric patients from Austria (153), France (110), Italy (97) and Germany (29), recruited in 1991–2010, 1999–2007, 1987–2010 and 1994–2008, respectively. In both cohorts, patients were treated according to a multimodal protocol consisting of multiagent chemotherapy mostly involving combinations of vincristine (V), ifosfamide (I), doxorubicin (D), cyclophosphamide (C), etoposide (E) and/or actinomycin-D (A), combined with surgery and/or radiation therapy. In the induction chemotherapy, three main protocols were used: VIDE (63% of patients), VDC (17%) and VDC + VAI + VDC + IE (14%) (supplementary Table S1, available at *Annals of Oncology* online). Postoperative chemotherapy typically consisted of the administration of the VAC or the VAI regimen. Of patients in the replication cohort with metastasis at diagnosis and/or non-resectable primary tumor, 18% were treated with high-dose chemotherapy. None of the patients in the discovery cohort received high-dose chemotherapy.

Relevant clinical information was abstracted from medical records (supplementary Table S2, available at *Annals of Oncology* online). Where possible, tumor response to treatment, defined as the percentage of necrosis induced in the tumor after neoadjuvant chemotherapy, was determined histologically. Overall survival (OS) was calculated as the time from tumor diagnosis until death from any cause or date last known to be alive.

Written informed consent was obtained from adult patients and from the parents or legal guardians of children. The study was approved by the ethics committees of all participating universities and hospitals.

candidate genes and SNP genotyping

We selected 24 genes reported to be involved in the pharmacokinetics of the six agents commonly used in chemotherapy regimen for ES, based on the information available in the database PharmaGKB [7] (supplementary Table S3, available at *Annals of Oncology* online). A total of 384 SNPs were selected across these candidate genes, as previously described [8].

Germline DNA, isolated from peripheral blood lymphocytes from participants in the discovery and replication cohorts, was genotyped using a customized Illumina GoldenGate VeraCode SNP genotyping assay (Illumina, San Diego, CA) on the BeadXpress platform according to the published protocol. Genotypes were called using GenomeStudio software. We excluded SNPs with a call rate <0.95 with minor allele frequency <0.05, whose genotype distribution deviated from the Hardy–Weinberg equilibrium ($P < 10^{-6}$), with Mendelian allele-transmission errors, or with discordant genotypes between duplicate samples. Samples with a call rate <0.90 were excluded.

statistical analysis

We studied the association of SNPs with tumor response and OS. Patients were divided into two categories: good responders, with tumor necrosis ≥90%; and poor responders, with tumor necrosis <90% [9]. Odds ratios and 95% confidence intervals (CIs) for good tumor response by genotype were estimated using logistic regression analysis. SNPs for which associations with $P < 0.05$ were observed were assessed in the European replication cohort.

We also tested associations between SNP genotypes and OS using the Cox regression analysis. SNPs with $P < 0.05$ in the discovery set were assessed in the replication cohort.

Clinical factors with associated $P < 0.05$ in univariable analyses (supplementary Table S2, available at *Annals of Oncology* online) with tumor response or OS were included as covariates in corresponding multivariable analyses.

In addition to the additive genetic model, we considered dominant and recessive models.

Analysis were carried out using PLINK [10] (v. 1.07) or SPSS software (v. 18.0; SPSS Inc., Chicago, IL).

functional annotations

We used information from the Encyclopedia of DNA Elements (ENCODE) [11] using custom tracks on the UCSC Genome browser [12] and HaploReg [13] to investigate whether the risk-associated SNPs or their correlated SNPs ($r^2 \geq 0.8$) had potential regulatory functions.

results

The demographic and clinical characteristics of both cohorts are shown in supplementary Table S4, available at *Annals of Oncology* online. After filtering, 334 SNPs of the 384 genotyped were successfully analyzed (Figure 1). There was no evidence of departure from the Hardy–Weinberg equilibrium for any. Data for two patients in the Spanish cohort and 55 patients in the replication cohort were excluded due to a low genotyping call rate (<0.90), leaving 104 and 334 patients, respectively.

associations with tumor response to treatment

Associations with tumor response were assessed in 77 patients from the Spanish discovery cohort for which this information was available. After adjusting for age and the presence of metastasis at diagnosis, an association with $P < 0.05$ was observed for 20 SNPs. However, none of these associations were replicated in the European cohort ($n = 197$, $P \geq 0.05$) (Figure 1).

Also including neoadjuvant therapy as an additional covariate in multivariable models made no substantial difference to the results obtained (data not shown).

associations with OS

Since adjustment for tumor response made no substantial difference to the estimated HR (based on an analysis of the cases for which this information was available), we present results without adjustment for this covariate, based on a larger sample size.

We identified 43 SNPs associated with OS at $P < 0.05$ in the Spanish cohort after adjusting for age at diagnosis, presence of metastasis at diagnosis and recurrence ($n = 97$) (supplementary Table S5, available at *Annals of Oncology* online). Associations with three of these were replicated in the European cohort ($n = 305$) (Table 1, supplementary Table S5 and Figure S1, available at *Annals of Oncology* online). The strongest evidence of association was found for the SNP rs7190447, an intronic polymorphism in *ABCC6*. In both cohorts, a recessive model was the best fit; C-allele homozygotes had a higher risk of death

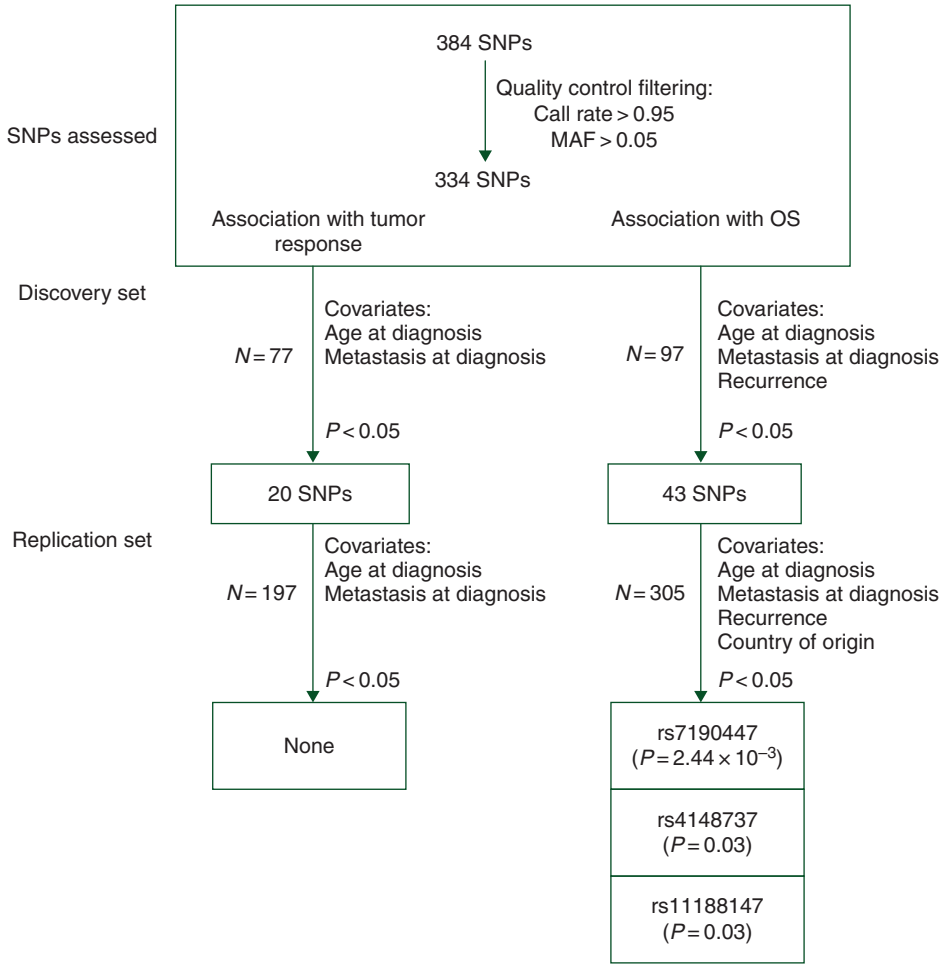


Figure 1. Flow chart of the study. SNPs, single-nucleotide polymorphisms; OS, overall survival; MAF, minor allele frequency.

(discovery phase: HR = 14.30, 95% CI = 1.53–134, $P = 0.020$; replication phase: HR = 9.28, 95% CI = 2.20–39.2, $P = 0.0024$, supplementary Figure S1A, available at *Annals of Oncology* online). The 5-year survival for patients carrying 1 or 2 copies of the G allele was 74% in the discovery cohort and 65% in the replication series, while no individual carrying the CC genotype lived for 5 years following diagnosis (Table 1).

The G allele of rs4148737, an intronic SNP in *ABCB1*, was associated with poorer OS under a recessive model (discovery phase: HR = 2.96, 95% CI = 1.08–8.10, $P = 0.034$; replication phase: HR = 1.60, 95% CI = 1.05–2.44, $P = 0.029$, supplementary Figure S1B, available at *Annals of Oncology* online). The estimated 5-year survival for cases with AA/AG and GG genotypes was 75% and 66%, respectively, in the discovery cohort, and 66% and 55%, respectively, in the European cohort (Table 1).

Finally, the minor T allele for an SNP located 2.7 kb downstream of the *CYP2C8* gene, rs11188147, was associated with an increased risk of death under a recessive model (discovery phase: HR = 2.49, 95% CI = 1.06–5.87, $P = 0.037$; replication phase: HR = 1.77, 95% CI = 1.06–2.96, $P = 0.030$, supplementary Figure S1C, available at *Annals of Oncology* online). The 5-year survival for CC/CT and TT carriers was 76% and 67%, respectively, in the discovery cohort, and 65% and 58%, respectively, in the replication series (Table 1).

functional annotations

Functional annotations of SNPs rs7190447, rs4148737 and rs11188147 and correlated variants ($r^2 \geq 0.8$) are shown in supplementary Table S6 and Figures S2–S4, available at *Annals of Oncology* online.

SNP rs7190447 was found to be in perfect linkage disequilibrium (LD) with rs7192303, an intronic polymorphism located 122 bp upstream. The genomic region containing rs7192303 is enriched with specific histone marks associated with transcribed regions in a hepatocellular carcinoma cell line, with weak enhancers in a skeletal muscle myoblast cell line, and with a DNase hypersensitive cluster in 123 different cell types (supplementary Table S6 and Figure S1, available at *Annals of Oncology* online). The strongest and most robust ChIP-seq signal is observed for CTCF binding in a large number of ENCODE cell lines (69) (supplementary Table S6 and Figure S2, available at *Annals of Oncology* online). Strong signals for cohesin subunits RAD21 and SMC3 were also observed. The genomic region containing rs7192303 also has the potential to form chromatin loops, through CTCF binding, with intronic regions of *ABCC6* and *ABCC1*, both located upstream, in a breast cancer cell line (MCF-7) (supplementary Figure S3, available at *Annals of Oncology* online). We explored expression quantitative trait loci using Genotype-Tissue Expression (GTEx) data, and found

Table 1. Associations between SNPs and OS in ES patients

Gene	Chr	SNP ID	Position[a]	Location	Model	MAF	Discovery ($n = 97$)						Replication ($n = 305$)				
							Genotype	n	5-year OS	P value	HR	95% CI	n	5-year OS	P value	HR	95% CI
ABCC6	16	rs7190447	16289126	Intronic	Recessive	0.07	GG/GC	96	74%				303	65%			
							CC	1	–				2	–			
							Per allele C			0.020	14.3	1.53–134			0.0024	9.28	2.20–39.2
ABCB1	7	rs4148737	87171152	Intronic	Recessive	0.43	AA/AG	80	75%				251	66%			
							GG	17	66%				54	55%			
							Per allele G			0.034	2.96	1.08–8.10			0.029	1.60	1.05–2.44
CYP2C8	10	rs11188147	96793820	2.7 kb downstream	Recessive	0.39	CC/CT	72	76%				263	65%			
							TT	25	67%				42	58%			
							Per allele T			0.037	2.49	1.06–5.87			0.030	1.77	1.06–2.96

Associations between SNPs and OS were assessed including important clinical covariates (age and metastasis at diagnosis and recurrence) in the Cox regression analyses. Three models of inheritance were evaluated. Only SNP associations with $P < 0.05$ in the discovery cohort and replicated in the European cohort are shown. HRs are per copy of the specified minor allele.

SNP, single-nucleotide polymorphisms; MAF, minor allele frequency; 5-year OS, 5-year overall survival; OS, overall survival; HR, hazard ratio; CI, confidence interval.

[a]Chromosome positions are based on Genome Reference Consortium Human Build 37 (GRCh37/hg19).

statistically significant differences in *ABCC6* gene expression by rs7192303 genotype in esophagus muscularis, liver and vagina ($P = 0.017$, 0.019 and 0.048, respectively) (data not shown). These findings suggest rs7192303 as the most plausible causal SNPs for the observed association with OS.

The intronic SNP rs4148737 resides in a weakly transcribed region, but also overlaps with a weak enhancer in GM12878 and in a RUNX3 ChIP-seq cluster in the same lymphoblastoid cell line. It was predicted to overlap with a DNAse hypersensitive region in a lymphoblastoid cell line and in cerebellar and hippocampal astrocytic cell lines, and to alter EBF, ERα-a, Hic1 regulatory motifs (supplementary Table S6 and Figure S4, available at *Annals of Oncology* online). None of the nine intronic SNPs in high LD ($r^2 \geq 0.84$) with this SNP had stronger functional evidence reported (supplementary Table S6, available at *Annals of Oncology* online). According to GTEx data, rs4148737 influences *ABCB1* expression in testis ($P = 0.022$), breast ($P = 0.033$) and muscle skeletal ($P = 0.038$).

No strong functional evidence was observed for SNP rs11188147 or for any of the 18 variants that are in high LD with it ($r^2 \geq 0.83$) (supplementary Table S6, available at *Annals of Oncology* online). No significant differences in *CYP2C8* expression were found for any tissues available in GTEx.

discussion

To date, most research has focused on polymorphisms, mainly SNPs, in candidate genes as potential biomarkers of toxicity and efficacy for anticancer agents; however, genetic variants in a single gene would explain only a small proportion of the interindividual variability observed in drug efficacy and toxicity. The evaluation of genetic variants in all the genes within a biological or pharmacological pathway may identify additional predictive and prognostic variants and thereby improve our ability to personalize treatment by predicting drug response [14]. We have assessed associations with treatment outcome in children diagnosed with ES for 334 SNPs in 24 genes involved in the pharmacokinetics of chemotherapeutic agents used in ES therapy. This is the first pharmacogenetic study carried out for this disease to consider an integrated pathway-based approach.

We have identified and replicated associations with OS for three variants: rs7190447, rs4148737 and rs11188147, located in *ABCC6*, *ABCB1* and *CYP2C8*.

Both *ABCC6* and *ABCB1* are members of the ATP-binding cassette (ABC) transporter superfamily, which have been implicated in mediating multidrug resistance in tumor cells reducing effectiveness of chemotherapeutics and decreasing survival [15]. *ABCC6*-transfected Chinese hamster ovary cells exhibited enhanced resistance to a variety of antineoplastics, including drugs administered to ES patients (etoposide, doxorubicin and actinomycin-D) [16]. In addition, *ABCC6* is expressed in tumors of patients with localized ES [17], so a role for *ABCC6* in multidrug resistance of ES cells is plausible. It remains to be determined whether intronic polymorphisms in *ABCC6* have an impact on gene expression and hence an effect on pharmacodynamics [15]. Recent studies have demonstrated that variants in non-transcribed regions can influence gene expression through regulatory mechanisms [18]. ENCODE and HaploReg data suggest that the genomic region containing the SNP rs7192303 (in perfect LD with our

replicated SNP, rs7190447) might be regulatory; SNP rs7192303 could affect CTCF binding to DNA and it is through CTCF that the overlapping genomic region could form a chromatin loop with upstream intronic regions of *ABCC6* and *ABCC1*. *CTCF* encodes an 11 zinc finger DNA-binding protein involved in diverse genomic regulatory functions, and it has recently been shown that *CTCF* regulates various aspects of gene expression and the establishment of genome topology by mediating long-range chromatin interactions [19]. We hypothesize that differences in *ABCC6* expression caused by changes in CTCF binding due to rs7192303 could affect the efflux of *ABCC6* target-drugs used in ES standard treatment, thus affecting intracellular drug levels in ES tumor cells, which ultimately determines the effectiveness of chemotherapy. On the other hand, the highest levels of *ABCC6* mRNA and protein expression have been detected in the liver and kidney [20] and, based on the GTEx data, the liver is one of the tissues where rs7192303 genotypes potentially influence *ABCC6* expression. ABC efflux pumps located in hepatocytes and kidney proximal tubule cells are crucial for drug elimination [21, 22]; while little is known about the physiological role of *ABCC6* [23], altered expression of this protein in hepatocytes may have an impact on the systemic bioavailability of drugs, and therefore on treatment response and patient survival.

To date, *ABCB1* and *ABCC1* are the only ABC genes that have been investigated in some detail in ES; however, the findings are contradictory. Although a significant association between protein expression and poorer response to therapy in pre- and post-therapeutic ES has been described [24], *ABCB1* mRNA and protein expression was not predictive of prognosis [25, 26]. Genetic polymorphisms in *ABCB1* have been reported to change mRNA/protein expression and function; however, little attention has been given to intronic and non-coding SNPs in this gene, and their possible link to cancer [27]. Consistent with a previous study in which we reported a significant association for the minor G allele of rs4148737 with poorer OS in pediatric osteosarcoma patients, the most common pediatric bone tumor (HR = 3.66, 95% CI = 1.85–6.11, $P = 6.9 \times 10^{-5}$) [8], in the current work, we observed that the GG genotype was associated with higher risk of death, suggesting that rs4148737 may be important as a prognostic marker after treatment in pediatric bone tumors.

CYP2C8 plays a role in the oxidative metabolism of some drugs used in ES treatment, in particular cyclophosphamide and ifosfamide [28]. Although it has been shown that there is great interindividual variation in the metabolism of *CYP2C8*-specific substrates and in *CYP2C8* expression [29], nothing has previously been reported about the impact of *CYP2C8* polymorphisms and their implications for clinical outcome in patients treated with cyclophosphamide and ifosfamide.

While one might expect the three replicated genetic variants also to be associated with tumor response, we did not observe this association. The evaluation of tumor response after administration of ES neoadjuvant therapy, when chemotherapy treatment had not been completed, could explain the lack of observed association with this clinical feature, particularly bearing in mind that *ABCB1* and *ABCC6* not only transport neoadjuvant drugs but also adjuvant chemotherapeutics, and *CYP2C8* is in part responsible for the oxidative metabolism of neoadjuvant and adjuvant ES agents. On the other hand, lack of replication could be ascribed to treatment heterogeneity between patients included in the two cohorts. We evaluated associations between SNP genotypes and tumor response, including neoadjuvant therapy as a covariate, in addition to age and metastasis at diagnosis, in order to assess the possibility that different neoadjuvant regimens could affect association with tumor response. Although we obtained the same significant associations in the discovery cohort, none were replicated in the European cohort (data not shown). Small sample size could be also a reason for the failure to replicate associations.

conclusion

We have identified genetic variants in the *ABCC6*, *ABCB1* and *CYP2C8* genes that were significantly associated with OS in ES patients. These findings highlight the clinical relevance of these genes as prognostic markers, although experimental verification of putative regulatory function will be required.

acknowledgements

We thank Javier Benitez for his comments on the manuscript and Daniela Caronia and Eva Sorz for data collation.

funding

This work was supported by the Spanish Association against Cancer (AECC: Asociación Española contra el Cáncer). Human Genotyping lab is a member of CeGen, PRB2-ISCIII and is supported by grant PT13/0001, of the PE I + D + i 2013–2016, funded by ISCIII (Instituto de Salud Carlos III) and FEDER (Fondo Europeo de Desarrollo Regional). This study was also supported by grants from the Ligue Nationale Contre le Cancer (Equipe labellisée), and the European PROVABES (ERA-649 NET TRANSCAN JTC-2011), ASSET (FP7-HEALTH-2010-259348), and EEC (Euro Ewing Consortium) (HEALTH-F2-2013-602856) projects. SR-P is a predoctoral fellow supported by the Severo Ochoa Excellence Programme (Project SEV-2011-0191). GG-M and JA are supported by Asociación Pablo Ugarte, Miguelañez S.A, ASION and Instituto de Salud Carlos III (PI12/00816 and RD12/0036/0027). KS is supported by grants from the Italian Association for Cancer Research–AIRC (CIG_14049) and by Italian Ministry of Health–TRANSCAN_Provabes and Piero Picci is supported by Italian Ministry of Health–TRANSCAN_Provabes.

disclosure

The authors have declared no conflicts of interest.

references

1. Esiashvili N, Goodman M, Marcus RB, Jr. Changes in incidence and survival of Ewing sarcoma patients over the past 3 decades: Surveillance Epidemiology and End Results data. J Pediatr Hematol Oncol 2010; 30: 425–430.
2. Evans WE, Relling MV. Pharmacogenomics: translating functional genomics into rational therapeutics. Science 1999; 286: 487–491.
3. Arpaci E, Yetisyigit T, Seker M et al. Prognostic factors and clinical outcome of patients with Ewing's sarcoma family of tumors in adults: multicentric study of the Anatolian Society of Medical Oncology. Med Oncol 2013; 30: 469.

4. Stahl M, Ranft A, Paulussen M et al. Risk of recurrence and survival after relapse in patients with Ewing sarcoma. Pediatr Blood Cancer 2011; 57: 549–553.
5. Sissung TM, Troutman SM, Campbell TJ et al. Transporter pharmacogenetics: transporter polymorphisms affect normal physiology, diseases, and pharmacotherapy. Discov Med 2012; 13: 19–34.
6. Ventola CL. Role of pharmacogenomic biomarkers in predicting and improving drug response: part 1: the clinical significance of pharmacogenetic variants. P T 2013; 38: 545–560.
7. The Pharmacogenomics Knowledge database. htpps://www.pharmgkb.com.
8. Caronia D, Patiño-Garcia A, Peréz-Martínez A et al. Effect of ABCB1 and ABCC3 polymorphisms on osteosarcoma survival after chemotherapy: a pharmacogenetic study. PLoS One 2011; 6: e26091.
9. Wunder JS, Paulian G, Huvos AG et al. The histological response to chemotherapy as a predictor of the oncological outcome of operative treatment of Ewing sarcoma. J Bone Joint Surg Am 1998; 80: 1020–1033.
10. Purcell S, Neale B, Todd-Brown K et al. PLINK: a toolset for whole-genome association and population-based linkage analysis. Am J Hum Genet 2007; 81: 559–575.
11. Myers RM, Stamatoyannopoulos J, Snyder M et al. A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol 2011; 9: e1001046.
12. UCSC Genome Browser. https://genome.ucsc.edu/cgi-bin/hgGateway.
13. Haploreg web interface. http://www.broadinstitute.org/mammals/haploreg/haploreg.php.
14. Ekhart C, Rodenhuis S, Smits PH et al. An overview of the relations between polymorphisms in drug metabolising enzymes and drug transporters and survival after cancer drug treatment. Cancer Treat Rev 2009; 35: 18–31.
15. Cascorbi I, Haenisch S. Pharmacogenetics of ATP-binding cassette transporters and clinical implications. Methods Mol Biol 2010; 596: 95–121.
16. Belinsky MG, Chen ZS, Shchaveleva I et al. Characterization of the drug resistance and transport properties of multidrug resistance protein 6 (MRP, ABCC6). Cancer Res 2002; 62: 6172–6177.
17. Svoboda LK, Harris A, Bailey NJ et al. Overexpression of HOX genes is prevalent in Ewing sarcoma and is associated with altered epigenetic regulation of developmental transcription programs. Epigenetics 2014; 9: 1613–1625.
18. Zhang X, Bailey SD, Lupien M. Laying a solid foundation for Manhattan—'setting the functional basis for the post-GWAS era'. Trends Genet 2014; 30: 140–149.
19. Ong CT, Corces VG. CTCF: an architectural protein bridging genome topology and function. Nat Rev Genet 2014; 15: 234–246.
20. Beck K, Hayashi K, Nishiguchi B et al. The distribution of Abcc6 in normal mouse tissues suggests multiple functions for this ABC transporter. J Histochem Cytochem 2003; 51: 887–902.
21. Köck K, Brouwer KL. A perspective on efflux transport proteins in the liver. Clin Pharmacol Ther 2012; 92: 599–612.
22. Masereeuw R, Russel FG. Regulatory pathways for ATP-binding cassette transport proteins in kidney proximal tubules. AAPS J 2012; 14: 883–894.
23. Le Saux O, Martin L, Aherrahrou Z et al. The molecular and physiological roles of ABCC6: more than meets the eye. Front Genet 2012; 3: 289.
24. Roessner A, Ueda Y, Bockhorn-Dworniczak B et al. Prognostic implication of immunodetection of P glycoprotein in Ewing's sarcoma. J Cancer Res Clin Oncol 1993; 119: 185–189.
25. Perri T, Fogel M, Mor S et al. Effect of P-glycoprotein expression on outcome in the Ewing family of tumors. Pediatr Hematol Oncol 2001; 18: 325–334.
26. Roundhill E, Burchill S. Membrane expression of MRP-1, but not MRP-1 splicing or Pgp expression, predicts survival in patients with ESFT. Br J Cancer 2013; 109: 195–206.
27. Fung KL, Gottesman MM. A synonymous polymorphism in a common MDR1 (ABCB1) haplotype shapes protein function. Biochim Biophys Acta 2009; 1794: 860–871.
28. Zanger UM, Schwab M. Cytochrome P450 enzymes in drug metabolism: regulation of gene expression, enzyme activities, and impact of genetic variation. Pharmacol Ther 2013; 138: 103–141.
29. Daily EB, Aquilante CL. Cytochrome P450 2C8 pharmacogenetics: a review of clinical studies. Pharmacogenomics 2009; 10: 1489–1510.

SUPPLEMENTARY FIGURE LEGENDS

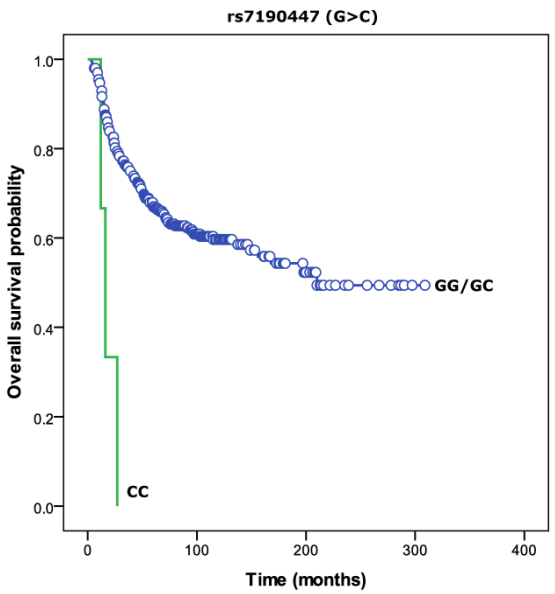
Supplementary Figure S1. Kaplan-Meier survival curves for ES patients (discovery and replication cohorts combined) according to genotype for (A) rs7190447 in *ABCC6* ($N_{GG/GC}$=399, N_{CC}=3, χ^2 = 14.84, $P_{log\text{-}rank}$=1.17×10^{-4}); (B) rs4148737 in *ABCB1* ($N_{AA/AG}$=326, N_{GG}=76, χ^2=3.73, $P_{log\text{-}rank}$=0.053); and (C) rs11188147 in *CYP2C8* ($N_{CC/CT}$=333, N_{TT}=69, χ^2 = 0.15, $P_{log\text{-}rank}$ = 0.69).

Supplementary Figure S2. ENCODE functional evidence displayed in the UCSC Genome Browser for rs7190447 and nearby SNPs. (A) Genomic location of the ABCC6 gene. Multiple DNase-seq and transcription factor ChIP-seq clusters can be observed. (B) Genomic location of rs7190447 (highlighted). A DNase hypersensitivity region was observed in 123 ENCODE cells around rs7190447 and overlapping with rs7192303 (highlighted), an intronic polymorphism located 122 pb upstream from rs7190447 and in perfect linkage disequilibrium (LD). Multiple transcription factors binding clusters can be observed in a large number of cells (identified by single-letter abbreviations). Gray boxes indicate the extent of the hypersensitive region for DNaseI hypersensitivity clusters or transcription factor occupancy, with the darkness of the box proportional to the maximum signal strength observed in any cells contributing to the cluster. The number to the left of a DNaseI hypersensitivity box shows how many cells are hypersensitive in the region. Within a ChIP-seq cluster, green highlighting indicates the highest scoring site of a Factorbook-identified canonical motif for the corresponding factor (http://genome.ucsc.edu/). Chromatin states characterized by combinations of histone marks are also shown in different human cell lines. Each chromatin state is associated with a different segment color. Blue, insulator; light green, weak transcribed; dark green, transcriptional transition; yellow, weak enhancer. The genomic region containing SNPs, rs7190447 and rs7192303 is also enriched for CTCF-mediated chromatin interactions in MCF-7 breast cancer cells. MCF-7 CTCF ChIA-PET interactions are shown as a density graph of signal enrichment based on aligned read density. Due to space limitations, only a subset of cells where a CTCF-ChIP-seq peak is detected and CTCF ChIA-PET interactions are shown. GM12878, lymphoblastoid cells; K1-hESC, embryonic stem cells; K562, erythrocytic leukemia cells; HepG2, hepatocellular carcinoma cells; HUVEC, umbilical vein endothelial cells; HMEC, mammary epithelial cells; HSMM, skeletal muscle muscle myoblast cells; NHEK, normal epidermal keratinocytes cells. Abbreviations: Txn: transcription.

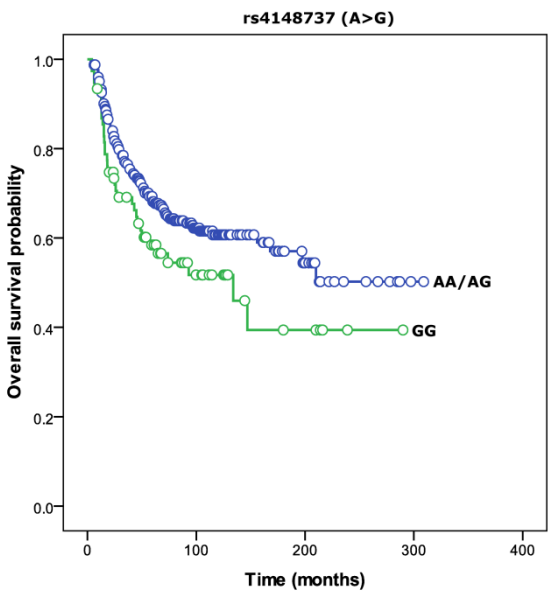
Supplementary Figure S3. CTCF-mediated chromatin interactions for rs7190447 and rs7192303 determined by chromatin interaction paired-end tag (ChiAPET) data from ENCODE. UCSC Genome Browser image of the genomic region containing rs7190447 (our replicated SNP) and rs7192303 (the SNP in perfect LD with it) showing ChIA-PET interactions and enrichment for CTCF in MCF-7 breast cancer cells. CTCF-mediated chromatin interactions are represented by two blocks, one at each end, connected by a horizontal line. The density graph shows the CTCF signal enrichment based on aligned read density. Not all MCF-7-CTCF ChIA-PET interactions are shown in full for the genomic region and the chromatin interactions have been adapted (red lines connecting blocks) to highlight relevant interactions only.

Supplementary Figure S4. ENCODE functional evidence displayed in the UCSC Genome Browser for rs4148737 and nearby SNPs. (A) Genomic location of the ABCB1 gene. Multiple DNase-seq and transcription factor ChIP-seq clusters can be observed. (B) Genomic location of rs4148737 (highlighted). Two DNase hypersensitivity regions are observed in five ENCODE cell lines around rs4148737, one of them overlapping. rs4148737 resides in a RUNX3 ChIP-seq cluster in lymphoblastoid cells (identified by G letter). Gray boxes indicate the extent of the hypersensitive region for DNaseI hypersensitivity clusters or transcription factor occupancy, with the darkness of the box proportional to the maximum signal strength observed in any cells contributing to the cluster. The number to the left of a DNaseI box shows how many cells are hypersensitive in the region. Within a ChIP-seq cluster, green highlighting indicates the highest scoring site of a Factorbook-identified canonical motif for the corresponding factor (http://genome.ucsc.edu/). Chromatin states characterized by combinations of histone marks are also shown in different human cell lines. Each chromatin state is associated with a different segment color. Yellow, weak enhancer; light green, weak transcribed. GM12878, lymphoblastoid cells; K1-hESC, embryonic stem cells; K562, erythrocytic leukemia cells; HepG2, hepatocellular carcinoma cells; HUVEC, umbilical vein endothelial cells; HMEC, mammary epithelial cells; HSMM, skeletal muscle muscle myoblast cells; NHEK, normal epidermal keratinocytes cells. Abbreviations: Txn: transcription.

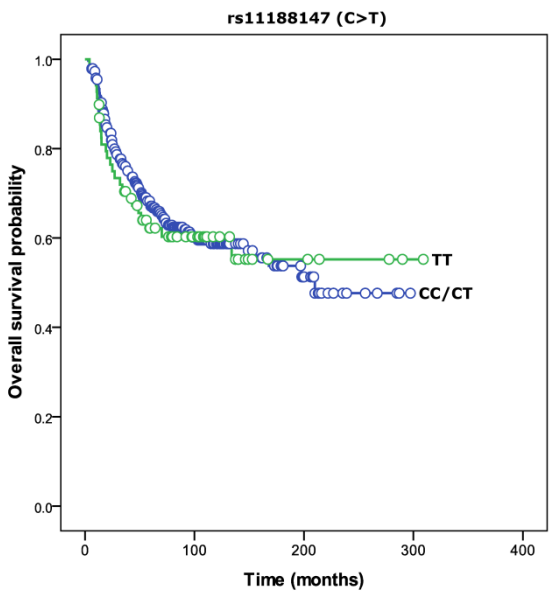
A)



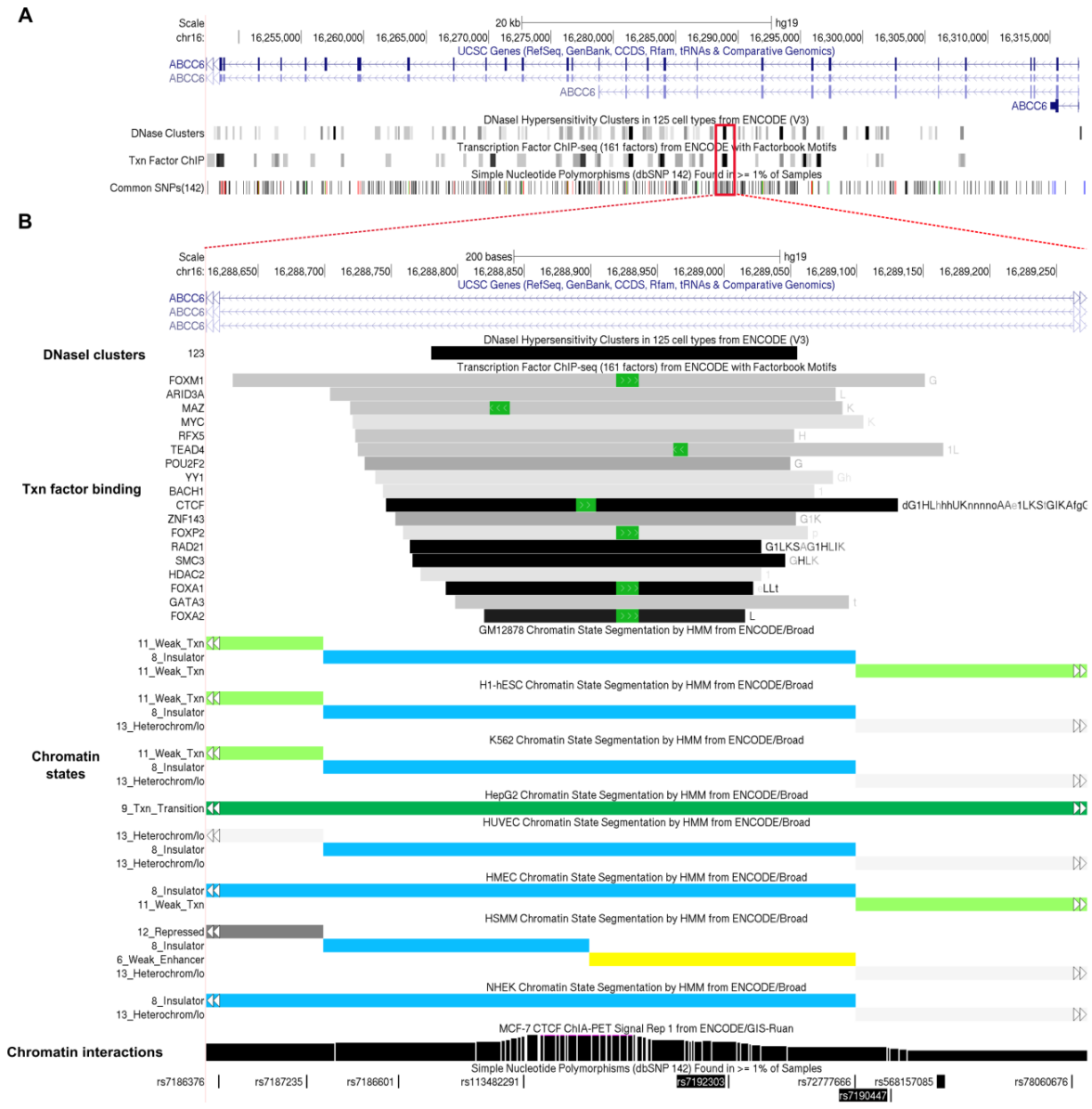
B)

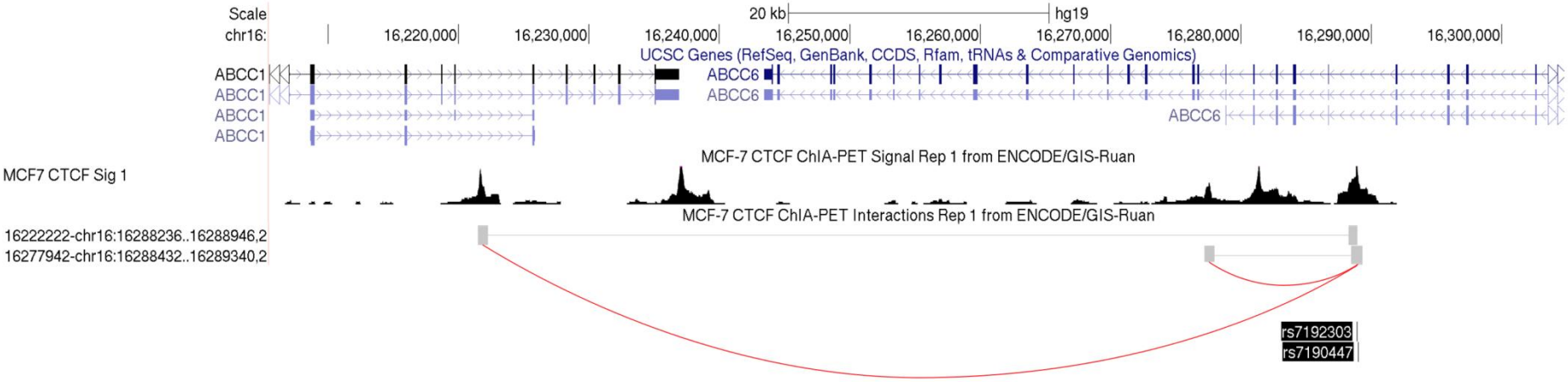


C)



Supplementary Figure S1


A
Scale
chr16:
20 kb
hg19
16,255,000
16,260,000
16,265,000
16,270,000
16,275,000
16,280,000
16,285,000
16,290,000
16,295,000
16,300,000
16,305,000
16,310,000
16,315,000
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
ABCC6
DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3)
DNase Clusters
Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs
Txn Factor ChIP
Simple Nucleotide Polymorphisms (dbSNP 142) Found in >= 1% of Samples
Common SNPs(142)
B
Scale
chr16:
200 bases
hg19
16,288,650
16,288,700
16,288,750
16,288,800
16,288,850
16,288,900
16,288,950
16,289,000
16,289,050
16,289,100
16,289,150
16,289,200
16,289,250
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
ABCC6
DNaseI clusters
DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3)
Txn factor binding
Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs
FOXM1
ARID3A
MAZ
MYC
RFX5
TEAD4
POU2F2
YY1
BACH1
CTCF
ZNF143
FOXP2
RAD21
SMC3
HDAC2
FOXA1
GATA3
FOXA2
Chromatin states
GM12878 Chromatin State Segmentation by HMM from ENCODE/Broad
11_Weak_Txn
8_Insulator
H1-hESC Chromatin State Segmentation by HMM from ENCODE/Broad
13_Heterochrom/lo
K562 Chromatin State Segmentation by HMM from ENCODE/Broad
9_Txn_Transition
HepG2 Chromatin State Segmentation by HMM from ENCODE/Broad
HUVEC Chromatin State Segmentation by HMM from ENCODE/Broad
HMEC Chromatin State Segmentation by HMM from ENCODE/Broad
HSMM Chromatin State Segmentation by HMM from ENCODE/Broad
12_Repressed
6_Weak_Enhancer
NHEK Chromatin State Segmentation by HMM from ENCODE/Broad
Chromatin interactions
MCF-7 CTCF ChIA-PET Signal Rep 1 from ENCODE/GIS-Ruan
Simple Nucleotide Polymorphisms (dbSNP 142) Found in >= 1% of Samples
rs7186376
rs7187235
rs7186601
rs113482291
rs7192306
rs72777666
rs568157085
rs78060676
rs7190447


Scale
20 kb
hg19
chr16:
16,220,000
16,230,000
16,240,000
16,250,000
16,260,000
16,270,000
16,280,000
16,290,000
16,300,000
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
ABCC1
ABCC1
ABCC1
ABCC1
ABCC6
ABCC6
ABCC6
MCF7 CTCF Sig 1
MCF-7 CTCF ChIA-PET Signal Rep 1 from ENCODE/GIS-Ruan
MCF-7 CTCF ChIA-PET Interactions Rep 1 from ENCODE/GIS-Ruan
16222222-chr16:16288236..16288946,2
16277942-chr16:16288432..16289340,2
rs7192303
rs7190447


A
Scale
chr7:
20 kb
hg19
87,140,000
87,150,000
87,160,000
87,170,000
87,180,000
87,190,000
87,200,000
87,210,000
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
ABCB1
ABCB1
DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3)
DNase Clusters
Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs
Txn Factor ChIP
Simple Nucleotide Polymorphisms (dbSNP 142) Found in >= 1% of Samples
Common SNPs(142)
B
Scale
chr7:
200 bases
hg19
87,170,950
87,171,000
87,171,050
87,171,100
87,171,150
87,171,200
87,171,250
87,171,300
87,171,350
87,171,400
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
ABCB1
ABCB1
DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3)
DNaseI clusters
2
3
Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs
Txn factor binding
RUNX3
G
GM12878 Chromatin State Segmentation by HMM from ENCODE/Broad
13_Heterochrom/lo
11_Weak_Txn
7_Weak_Enhancer
H1-hESC Chromatin State Segmentation by HMM from ENCODE/Broad
13_Heterochrom/lo
K562 Chromatin State Segmentation by HMM from ENCODE/Broad
13_Heterochrom/lo
HepG2 Chromatin State Segmentation by HMM from ENCODE/Broad
Chromatin states
11_Weak_Txn
HUVEC Chromatin State Segmentation by HMM from ENCODE/Broad
13_Heterochrom/lo
HMEC Chromatin State Segmentation by HMM from ENCODE/Broad
13_Heterochrom/lo
HSMM Chromatin State Segmentation by HMM from ENCODE/Broad
13_Heterochrom/lo
NHEK Chromatin State Segmentation by HMM from ENCODE/Broad
13_Heterochrom/lo
Simple Nucleotide Polymorphisms (dbSNP 142) Found in >= 1% of Samples
rs28381937
rs28381936
rs4148737
rs4148736

Supplementary Table S1. Neoadjuvant treatment given to ES patients					
Neoadjuvant chemotherapy	Discovery (N=106)	Replication (N=389)			
	Spain (106) N (%)	Germany (29) N (%)	Italy (97) N (%)	Austria (153) N (%)	France (110) N (%)
VIDE	27 (25%)	15 (52%)	-	153 (100%)	110 (100%)
VDC	79 (75%)	-	3 (3%)	-	-
VDC+VAI+VDC+IE	-	-	70 (72%)	-	-
VDI+C*E*+VDI+C*E	-	-	15 (15%)	-	-
VDCA+I	-	-	5 (5%)	-	-
VDIA	-	4 (14%)	-	-	-
EVDIA	-	4 (14%)	-	-	-
VDIA+EVDIA		1 (3%)	-	-	-
VIDE+VAI	-	3 (10%)	-	-	-
VIDE+VAI+VAC	-	1 (3%)	-	-	-
Other	-	1 (3%)	-	-	-
Missing	-	-	4 (4%)	-	-

Abbreviations: V, vincristine; I; ifosfamide; D, doxorubicin; E, etoposide; E* high-dose etoposide, A, actinomycin-D; C, cyclophosphamide; C* high-dose cyclophosphamide. Other: chemotherapy regimen involving vincristine, doxorubicin, etoposide and cisplatin.

Supplementary Table S2. Clinical information recorded from ES patients and associations in univariable analyses with tumor response and OS

	Tumor response		OS	
	Discovery	Replication	Discovery	Replication
Clinical information	*P*	*P*	*P*	*P*
Age at diagnosis	0.039	0.002	0.048	0.038
Gender	0.52	0.38	0.55	0.79
Primary tumor site	0.29	0.99	0.33	0.69
Metastasis at diagnosis	0.001	0.042	$1.67x10^{-6}$	$6.8x10^{-12}$
Neoadjuvant therapy	0.55	0.54	0.15	0.15
Response to treatment	NA	NA	0.029	$6.0x10^{-5}$
Recurrence*	0.093	0.082	$1.67x10^{-6}$	$8.7x10^{-22}$
Vital status	NA	NA	NA	NA
Overall survival (OS)	NA	NA	NA	NA
Country of origin	NA	0.90	NA	0.81

Information on age at diagnosis, sex, primary tumor site, existence of metastasis at diagnosis, tumor response, treatment protocol, vital status and development of recurrence was abstracted retrospectively from medical records. *Recurrence was defined as any evidence of new disease during/after the completion of therapy, including both locoregional and distant disease relapses. Associations between clinical factors were assessed in univariable analyses with tumor response by logistic regression analyses and with OS by Cox regression analyses in the discovery and replication cohorts. Clinical factors with associated $P<0.05$ (in bold) in univariable analyses with tumor response or OS were included as covariates in corresponding multivariable analyses. Country of origin was included as covariate in analyses with OS in the replication cohort due to incomplete information regarding adjuvant therapy protocols. Abbreviations: OS, overall survival; NA, not applicable

Supplementary Table S3. Candidates genes studied	
Category	**Genes**
Transporters	*ABCA3, ABCB1, ABCC1, ABCC2, ABCC3, ABCC4, ABCC6, ABCG2, SLC31A1, SLCO6A1, SLC19A1*
Phase I metabolism enzymes	*MPO, SOD1, ALDH1A1, CYP3A4, CYP3A5, CYP2A6, CYP2B6, CYP2C8, CYP2C9, CYP2C19*
Phase II metabolism enzymes	*GSTM1, GSTP1, GSTT1*
24 candidate genes reported to be involved in the pharmacokinetics of the 6 agents (vincristine, ifosfamide, doxorubicin, cyclophosphamide, etoposide and actinomycin-D) commonly used in chemotherapy regimens for ES, were selected based on the information available in the database PharmaGKB	

Supplementary Table S4. Clinical characteristics of ES patients				
	Discovery (N=106)		Replication (N=389)	
Characteristic	**N**	**%***	**N**	**%***
Age at diagnosis (years)				
Median	12.2		14.5	
Range	0.4-27.8		0.1-27.8	
Sex				
Female	42	39.6	156	40.1
Male	64	60.4	233	59.9
Primary site				
Upper extremities	9	8.7	39	10.2
Lower extremities	48	46.2	129	33.9
Axial	41	39.4	203	53.3
Soft tissue	6	5.8	10	2.6
Missing	2		8	
Metastasis at diagnosis				
No	62	59.6	255	67.3
Yes	42	40.4	124	32.7
Missing	2		10	
Response to treatment				
Good	55	70.5	148	64.3
Poor	23	29.5	82	35.7
Missing	28		159	
Relapse				
No	67	67.0	211	58.1
Yes	33	33.0	152	41.9
Missing	6		26	
Vital status				
Alive	67	67.0	229	61.7
Dead	33	33.0	142	38.3
Missing	6		18	
Follow-up (years)				
Median	93.7		70.3	
Range	7.8-300		12.3-312	

* Percentages are computed based on the total number of non-missing values*Percentages are computed based on the total number of non-missing values.

Supplementary Table S5. Analysis of associations between SNPs and OS											
Gene	Chr.	Variant	Position *	Location	Model	Discovery (N=97)			Replication (N=305)		
						P	Minor allele HR	95%CI	*P*	Minor allele HR	95%CI
ABCC1	16	rs212081	16225971	Intronic	Recessive	$2.59x10^{-4}$	T 6.62	2.40-18.3	0.46	T 1.19	0.75-1.90
SLCO6A1	5	rs981988	101819981	Intronic	Recessive	$1.52x10^{-3}$	G 54.2	4.59-639	0.11	G 0.20	0.03-1.41
GSTP1	11	rs7927381	67346743	4.3kb upstream	Recessive	$1.52x10^{-3}$	T 54.2	4.59-639	0.28	T 3.06	0.39-23.8
ABCB1	7	rs7787082	87157051	Intronic	Additive	$1.91x10^{-3}$	A 0.29	0.13-0.64	0.90	A 1.00	0.71-1.42
ABCC3	17	rs12451302	48751164	Intronic	Recessive	$2.62x10^{-3}$	G 5.16	1.77-15.0	0.29	G 0.80	0.52-1.22
CYP2C9	10	rs4918758	96697252	2kb upstream	Recessive	$2.98x10^{-3}$	C 5.11	1.74-15.0	0.09	C 1.55	0.93-2.59
ABCC3	17	rs3785912	48756937	Intronic	Recessive	$4.06x10^{-3}$	A 5.99	1.77-20.3	0.25	A 0.70	0.38-1.29
ABCC1	16	rs3888565	16183045	Intronic	Additive	$4.69x10^{-3}$	A 0.28	0.12-0.68	0.25	A 0.83	0.59-1.15
CYP3A5	*7*	rs28365067	99272310	Intronic	Additive	$5.84x10^{-3}$	T 4.42	1.54-12.7	0.89	T 1.04	0.62-1.74
ABCC1	16	rs35621	16168608	Intronic	Additive	$7.27x10^{-3}$	T 0.21	0.07-0.66	0.97	T 0.99	0.68-1.46
ABCB1	7	rs2214102	87229501	5´ UTR	Recessive	$8.15x10^{-3}$	A 28.4	2.38-339	-	-	-
ABCC1	16	rs16967755	16199255	Intronic	Additive	0.01	G 0.31	0.13-0.76	0.64	G 1.07	0.80-1.45
ABCC6	16	rs16967488	16252696	Intronic	Dominant	0.01	C 2.80	1.24-6.35	0.73	C 1.07	0.74-1.53

ALDH1A1	9	rs348481	75514436	1.1kb downstream	Recessive	0.01	C 8.03	1.53-42.0	0.99	C 0.99	0.31-3.15
ABCC2	10	rs717620	101542578	5´ UTR	Recessive	0.02	A 15.8	1.72-146	0.91	A 0.93	0.23-3.77
ABCC1	16	rs35626	16170615	Intronic	Additive	0.02	T 0.41	0.20-0.85	0.90	T 0.98	0.74-1.30
CYP2C9	10	rs11597626	96604273	Intronic	Dominant	0.02	G 0.37	0.17-0.84	0.89	G 1.03	0.71-1.48
CYP2C9	10	rs12251688	96693727	4.6kb upstream	Dominant	0.02	T 0.38	0.17-0.84	0.83	T 1.04	0.72-1.49
CYP3A4	7	rs4646437	99365083	Intronic	Dominant	0.02	T 3.06	1.20-7.79	0.43	T 1.19	0.77-1.83
ABCC6	**16**	**rs7190447**	**16289126**	**Intronic**	**Recessive**	**0.02**	**C 14.30**	**1.53-134**	**2.44×10^{-3}**	**C 9.28**	**2.20-39.2**

Supplementary Table S5. Analysis of associations between SNPs and OS (continued)

Gene	Chr.	Variant	Position *	Location	Model	Discovery (N=97)			Replication (N=305)		
						P	Minor allele HR	95%CI	*P*	Minor allele HR	95%CI
ABCC1	16	rs12922404	16060994	Intronic	Dominant	0.02	T 2.61	1.16-5.89	0.17	T 1.29	0.89-1.88
ABCB1	7	rs2235048	87138511	Intronic	Recessive	0.02	C 2.88	1.16-7.12	0.46	C 0.85	0.55-1.31
ABCB1	7	rs17064	87133470	3´ UTR	Additive	0.02	T 4.08	1.22-13.7	0.96	T 0.99	0.60-1.61
ABCC6	16	rs2238469	16283071	Intronic	Recessive	0.03	A 4.36	1.17-16.3	0.15	A 0.23	0.03-1.73
ABCC3	17	rs8079432	48749883	Intronic	Additive	0.03	G 2.77	1.11-6.91	0.68	G 0.91	0.57-1.45
ABCC4	13	rs9590220	95906694	Intronic	Additive	0.03	T 0.38	0.16-0.91	0.13	T 1.28	0.93-1.75
ABCB1	7	rs10264990	87202615	Intronic	Recessive	0.03	C 3.21	1.11-9.26	0.10	C 1.54	0.92-2.59
GSTP1	11	rs614080	67347287	3.8kb upstream	Recessive	0.03	G 2.56	1.09-6.04	0.52	G 0.85	0.52-1.39
ABCC1	16	rs212087	16230290	Intronic	Recessive	0.03	T 0.34	0.13-0.91	0.10	T 1.46	0.93-2.30
ABCB1	**7**	**rs4148737**	**87171152**	**Intronic**	**Recessive**	**0.03**	**G 2.96**	**1.08-8.10**	**0.03**	**G 1.60**	**1.05-2.44**
ABCG2	4	rs2725264	89026109	Intronic	Additive	0.04	G 2.60	1.06-6.36	0.41	G 0.83	0.53-1.29
CYP1B1	2	rs4646429	38306935	3.6kb upstream	Additive	0.04	A 0.44	0.20-0.95	0.95	A 0.99	0.73-1.34
CYP2C8	**10**	**rs11188147**	**96793820**	**2.7kb downstream**	**Recessive**	**0.04**	**T 2.49**	**1.06-5.87**	**0.03**	**T 1.77**	**1.06-2.96**

MPO	17	rs7208693	56357818	Missense	Additive	0.04	A 2.17	1.04-4.51	0.76	A 0.93	0.54-1.47
ABCC1	16	rs4148354	16174506	Intronic	Dominant	0.04	G 0.42	0.19-0.96	0.87	G 1.07	0.69-1.57
CYP2C8	10	rs1934956	96828160	Intronic	Recessive	0.04	T 4.95	1.06-23.2	0.47	T 1.44	0.53-3.96
CYP2A6	19	rs8192729	41350996	Intronic	Additive	0.05	A 2.99	1.02-8.75	0.92	A 1.03	0.62-1.69
ABCC1	16	rs2299670	16220858	Intronic	Additive	0.05	G 0.51	0.26-0.99	0.69	G 1.06	0.79-1.43
SOD1	21	rs2070424	33039320	Intronic	Additive	0.05	G 3.22	1.02-10.2	0.06	G 1.54	0.98-2.41
LPO	17	rs8178407	56344656	Intronic	Dominant	0.05	G 2.61	1.01-6.77	0.77	G 0.95	0.66-1.37
ABCC1	16	rs11075295	16177687	Intronic	Recessive	0.05	G 9.14	1.02-82.1	0.96	G 0.98	0.36-2.65
ABCB1	7	rs13237132	87191669	Intronic	Additive	0.05	G 1.80	1.00-3.22	0.20	G 1.19	0.91-1.54
ABCC4	13	rs9590211	95892414	Intronic	Dominant	0.05	A 2.14	1.00-4.57	0.71	A 0.93	0.63-1.37

Associations between SNPs and OS in ES patients were assessed by the Cox regression analyses, adjusted for age and metastasis at diagnosis and recurrence. Only SNP associations with $P < 0.05$ in the discovery cohort were considered in the replication cohort. Three models of inheritance were evaluated. HRs are per copy of the specified minor allele. Variants shown in bold were those associated with overall survival in both cohorts at $P<0.05$. * Chromosome positions are based on Genome Reference Consortium Human Build 37 (GRCh37/hg19). Abbreviations: Chr, chromosome; SNP, single nucleotide polymorphisms; OS, overall survival; HR, hazard ratio; CI, confidence interval.

Supplementary Table S6. Analysis of functional annotations of rs7190447, rs4148737 and rs11188147 and correlated variants ($r^2 \geq 0.8$)						
	Correlated variant	**LD (r^2)**	**Enhancer histone marks**	**DNase**	**Proteins bound**	**Motifs changed**
rs7190447	-	-	-	HA-sp	CTCF	GATA, Pou3f2
rs7206048	rs7190447	0.96	K562	GM12892	MAFK	Egr-1,GATA
rs6498619	rs7190447	1	K562	-	-	Pax-5
rs8044613	rs7190447	1	-	-	-	Evi-1,Pou2f2
rs11862259	rs7190447	1	-	-	-	Mef2, NF-AT1
rs7184822	rs7190447	0.98	-	-	-	22 altered motifs
rs7186376	rs7190447	1	-	-	-	ERalpha-a, HNF4, TLX1, NFIC
rs7187235	rs7190447	1	-	-	-	AP-1
rs7186601	rs7190447	1	-	-	-	BDP1_disc2, GCNF, Nr2f2, p300_known1
rs7192303	rs7190447	1	HSMM	PanIsletD, AG09309, AG10803, HA-h, HA-sp, HGF, HIPEpiC, HNPCEpiC, HPdLF, HVMF	CTCF, SMC3, ZNF143, RAD21, FOXA1,FOXA2, GATA3	INSM1

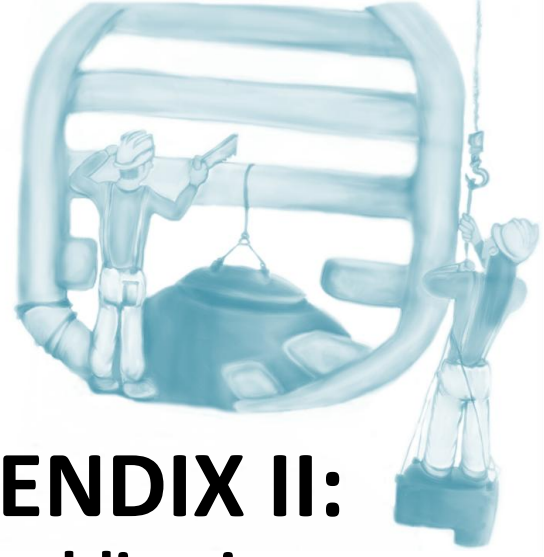
rs7199104	rs7190447	1	-	-	-	Foxa,Foxj2, Osf2
rs4148737	-	-	GM12878	GM12865,HA-h,HAc	-	EBF, ERalpha-a, Hic1
rs35572298	rs4148737	0.84	-	-	-	8 altered motifs
rs35280822	rs4148737	0.89	-	-	-	Homez,Lhx3
rs12154941	rs4148737	0.91	-	-	-	AP-1,Zfx
rs4148736	rs4148737	1	-	-	-	GR,Nanog
rs6961419	rs4148737	1	-	HConF,HFF-Myc,NHDF-neo	-	-
rs6961882	rs4148737	1	-	-	-	6 altered motifs
rs4148735	rs4148737	1	GM12878	Melano	-	GR, p300
rs1922242	rs4148737	1	-	-	-	5 altered motifs

Continue on next page

Supplementary Table S6. Analysis of functional annotations of rs7190447, rs4148737 and rs11188147 and correlated variants (r^2≥0.8) (continued)						
Variant	**Correlated variant**	**LD (r^2)**	**Enhancer histone marks**	**DNase**	**Proteins bound**	**Motifs changed**
rs2091766	rs4148737	0.88	-	-	-	8 altered motifs
rs11188147	-	-	-	-	-	-
rs1578436	rs11188147	1	-	-	-	9 altered motifs
rs7073968	rs11188147	1	-	-	-	HNF4, NF-I
rs10882517	rs11188147	1	-	-	-	Foxa, GZF1, Pou1f1
rs11188149	rs11188147	0.99	-	-	-	Pax-4
rs947173	rs11188147	0.99	-	-	-	BCL, BDP1, NRSF
rs1891070	rs11188147	0.99	-	-	-	Foxp1, Hdx
rs11572133	rs11188147	0.99	-	-	-	NRSF, Sin3Ak-20
rs12773510	rs11188147	0.99	-	-	-	11 altered motifs
rs199539470	rs11188147	0.95	-	-	-	10 altered motifs
rs58385086	rs11188147	0.99	-	-	-	11 altered motifs
rs11188156	rs11188147	0.98	-	-	-	DMRT3,Gfi1b

rs10882521	rs11188147	0.93	-	-	-	4 altered motifs
rs9702453	rs11188147	0.83	-	-	-	HNF4, Pdx1
rs145809484	rs11188147	0.90	-	-	-	GCM, Gcm1
rs143042734	rs11188147	0.88	-	-	-	5 altered motifs
rs13313110	rs11188147	0.98	-	-	-	Gfi1,Hsf,TATA
rs3752988	rs11188147	0.99	-	-	-	Ik-2,Mef2
rs10882525	rs11188147	0.99	-	-	-	4 altered motifs

The replicated variants associated with overall survival (in bold) and SNVs in strong (r^2>0.8) linkage disequilibrium (LD) with the replicated variants were analyzed using HaploReg to explore if they affected chromatin states or altered regulatory motifs or binding sites. AG09309, adult toe fibroblast cells; AG10803, abdominal skin fibroblast cells; GM12865, lymphoblastoid cells; GM12878, lymphoblastoid cells; GM12892, lymphoblastoid cells; HAc, human cerebellar astrocytic cells; HA-h, human hippocampal astrocytic cells; HA-sp, human spinal cord astrocytic cells; HConF, conjunctival fibroblast cells; HFF-Myc, foreskin fibroblast cells expressing canine cMyc; HGF, gingival fibroblasts cells; HIPEpiC, iris pigment epithelial cells; HNPCEpiC, non-pigment ciliary epithelial cells; HPdLF, periodontal ligament fibroblasts cells; HSMM, skeletal muscle myoblast cells; HVMF, villous mesenchymal fibroblast cells; K562, erythrocytic leukemia cells; Melano, human epidermal melanocyte cells; NHDF-neo,neonatal dermal fibroblast cells; PanIsletD, dedifferentiated human pancreatic islets



APPENDIX II:
other publications

- Apellaniz-Ruiz M*, Gallego C*, **Ruiz-Pinto S***, Carracedo A, Rodríguez-Antona C. Human Genetics: International Projects and Personalized Medicine. *Drug Metabolism and Personalized Therapy.* **31**, 3–8 (2016)

- **Ruiz-Pinto S***, Caronia D*, Martin M*, de la Torre J, Pita G, Moreno LT, Sastre J, Benítez J, García-Sáenz JA, González-Neira A. Predictive genetic markers of response for neoadjuvant doxorubicin versus docetaxel in primary breast cancer patients: a pharmacogenetic analysis of the NCT 00123929 phase 2 randomised trial. Under review in *Annals of Oncology*.